

Declaration of financial interests

Introducing a new policy for authors of research papers in *Nature* and *Nature* journals.

In the interests of transparency and to help readers to form their own judgements of possible bias, *Nature* and the *Nature* research journals will soon be encouraging authors to declare any competing financial interests in relation to research papers. For manuscripts with a received date on or after 1 October 2001, we will ask authors to fill out a form declaring any competing financial interests, shortly before final acceptance of the paper. We will publish a shortened version of this declaration as part of the paper, with a more detailed version, if appropriate, on our website. Authors may decline to disclose their financial interests, but we will publish the fact that they have declined to respond.

Hitherto, the policy of the *Nature* journals was that no declarations of competing interests were required from authors, and that potential referees should disqualify themselves from refereeing if they felt they had such a conflict (see *Nature* **385**, 469; 1997). The new policy is not based on any assumption that commercial interests of researchers are likely to lead to a lack of research integrity. Rather, it is based on a recognition of potential problems. There are three main reasons for adopting the policy.

- There is suggestive evidence in the literature that publication practices in biomedical research have been influenced by the commercial interests of authors. Several related discussions that contain the relevant references can be found in the *Journal of the American Medical Association*, 1 November 2000 (for example, A. D. DeAngelis Conflict of interest and the public trust. *J. Am. Med. Assoc.* **284**, 2237–2238; 2000; see also *Nature Neuroscience* **3**, 299; 2000). This evidence is consistent with the truism that although, in principle, science may be objective and its findings independent of other interests, scientists can be imperfect and subjective. There are circumstances in which selection of evidence, interpretation of results or emphasis of presentation might be inadvertently or even deliberately biased by a researcher's other interests.

- There is a more general concern among researchers and others about the possible undermining of the integrity of scientific research by increasing commercial links and consequent influences. We believe that the best way to maintain readers' trust in the integrity of the research we publish is through a policy of transparency. If financial interests are disclosed, readers will be able to make an informed judgement about their significance or lack of significance. We believe this will benefit both readers and authors alike.

- Many institutions have introduced policies on competing interests that require authors to include descriptions of financial and other interests in publications. We are happy to support them.

We do not expect to police this policy ourselves; we believe that primary responsibility for ensuring that researchers' conduct is appropriate lies with their employers, rather than with journal editors. However, where we believe trust has been significantly compromised by an author's actions, we will seek to redress the matter by an appropriate combination of sanctions and communications to readers and employers.

Authors will be invited to disclose any competing financial interests, using a standard declaration form to be completed by the corresponding author. Alternatively, authors may declare that they have

no competing financial interests. These declarations will be considered confidential before publication, and will not be disclosed to referees.

Authors who do not wish to make any financial disclosure will be asked to indicate as much on the form. This will in no way prejudice publication of their manuscript, but the fact that the authors have not made a declaration will be recorded, and will be distinguished from declarations of no competing financial interests.

For the purposes of this statement, competing interests are defined as those of a financial nature that, through their potential influence on behaviour or content or from perception of such potential influences, could undermine the objectivity, integrity or perceived value of a publication. They may include any of the following:

- **Funding:** Research support (including salaries, equipment, supplies, reimbursement for attending symposia, and other expenses) by organizations that may gain or lose financially through publication of the paper.

- **Employment:** Recent (that is, while engaged in the research project), present or anticipated employment by any organization that may gain or lose financially through publication of the paper.

- **Personal financial interests:** Stocks or shares in companies that may gain or lose financially through publication; consultation fees or other forms of remuneration from organizations that may gain or lose financially; patents or patent applications whose value may be affected by publication.

It is difficult to specify a threshold at which a financial interest becomes significant, although we note that many US universities require faculty members to disclose interests exceeding \$10,000 or 5% equity in a company (see, for example, B. Lo *et al.* *New Engl. J. Med.* **343**, 1616–1620; 2000). Any such figure is necessarily arbitrary, however, so we offer as one possible practical alternative guideline: "Any undeclared competing financial interests that could embarrass you were they to become publicly known after your work was published." We do not consider diversified mutual funds or investment trusts to constitute a competing financial interest. We will not require authors to state the monetary value of their financial interests.

For papers with more than one author, the author responsible for communication with the journal should provide a declaration on behalf of all authors. If the number of authors is so large that itemized disclosure is unfeasible, authors may choose to use the standard wording indicating as much.

We will continue to ask referees to exclude themselves as referees in cases where there is a significant conflict of interest, financial or otherwise. However, just as financial interests need not invalidate the conclusions of a paper, nor do they automatically disqualify a referee from evaluating it. We will ask referees to inform the editors of any related interests, including financial interests as defined above, that might be perceived as relevant. Editors will consider these statements when weighing referees' recommendations.

We welcome comments and suggestions about this policy, which should be sent to nature@nature.com, marked "Competing interests policy".

Philip Campbell PhD, Editor, *Nature*; Editor-in-Chief, *Nature Publications*.





In the bank
Japan faces struggle
to amass world-class
tissue collection
p754



Brain wave
Max Planck Society
plans centre for
neural imaging
p755



Dry run
Ecologists hope to
limit impact of
vanishing lake
p756



Collective action
Smithsonian staff
seek to influence
museum reforms
p757

Stem-cell giveaway proposed as confusion reigns over cell count

Jonathan Knight, San Francisco

Academic researchers in the United States should soon be able to access certain human embryonic stem-cell lines for free.

BresaGen, a company based in Adelaide, Australia, and ES Cell International (ESI) in Singapore say they plan to give their cells away to publicly funded scientists.

In return, the companies will ask for first rights to license any therapeutics and medical procedures that emerge from the research. The companies have developed ten human embryonic stem-cell lines between them.

The offer contrasts with the \$5,000 flat fee charged by the Wisconsin Alumni Research Foundation (WARF), which holds the US patent to human embryonic stem cells. WARF's private, non-profit laboratory says its fee is less than the cost of producing the cells.

Representatives from BresaGen will this week meet with officials in Washington to discuss how the company will distribute its cells to researchers funded by the National Institutes of Health (NIH). Meetings

between the NIH and the holders of other stem-cell lines are also being planned.

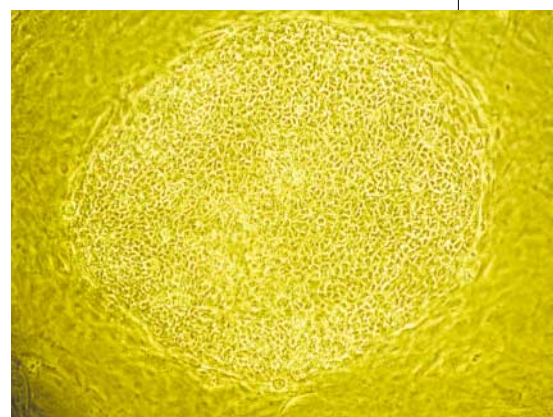
Allan Robins, BresaGen's chief scientist, says that one option being considered is a single, independent distribution centre that would handle all the cell lines for which federal funding will be permitted under rules announced by President George W. Bush on 9 August.

Such a centre could include all 60 of the stem-cell lines that the NIH says exist — 45 or so of which are not accounted for by WARF, ESI or BresaGen. The sources of these cells have not yet been identified, and many scientists are sceptical about their existence.



Cell for nothing:
BresaGen's chief
scientist Allan
Robins.

On 17 August, the American Association for the Advancement of Science (AAAS), one of the largest US scientific soci-



Sharing culture: BresaGen's stem cells could soon be available free of charge to researchers.

eties, issued a strongly worded statement calling on the Bush administration to publish the source of these stem-cell lines immediately.

Noting that Bush's policy depends heavily on the existence of the cell lines, the statement said that "it is essential that confusion over the actual number available be resolved as soon as possible". It also raised questions about whether cell lines derived abroad would meet US ethical standards. As of 20 August, the White House had made no response to the statement.

BresaGen has four of the known human embryonic stem-cell lines, which it derived at its US laboratory in Athens, Georgia. Robins says the company hopes to provide the cells to researchers free of charge. "We would like to be giving the cell lines away so they can be used as much as possible," he says.

ESI says it has already agreed to provide two of its six embryonic stem-cell lines to 15 laboratories worldwide, including several in the United States, for the cost of shipping.

ESI's lines were developed a year ago by a team of researchers headed by Alan Trounson at Monash University in Victoria, Australia (see *Nature Biotech.* **18**, 399–404; 2000).

"We are getting two or three requests a day since the Bush announcement," says ESI operations manager Catriona King.

Despite researchers' fears about the terms of access to stem-cell lines, the companies that have derived them appear keen to get

Partners turn to court over stem-cell rights

Within days of President Bush's announcement on the funding of stem-cell research, the first of many likely legal battles over stem-cell technology rights was underway.

On 13 August, just four days after Bush's statement, the Wisconsin Alumni Research Foundation (WARF) filed a lawsuit in an effort to keep its private partner Geron of Menlo Park, California, from taking too large a piece of the commercial pie. WARF holds the patent on human embryonic stem cells in the United States.

But shortly afterwards, both parties issued a statement saying that, despite the lawsuit, they were still negotiating with each

other on the patent, and that they expect to resolve the matter "in the near future".

Geron sponsored the research of James Thomson at the University of Wisconsin in Madison, who in 1998 derived the first human embryonic stem cells. In return, WARF, through which many University of Wisconsin inventions are patented, granted Geron limited rights to the technology.

Under the agreement, Geron has the right to commercialize any applications involving six of the more than 200 cell types in the human body — liver, muscle, nerve, bone, blood and the insulin-producing pancreatic islet cells.

At the end of July, Geron

exercised an option in the contract to claim an additional 12 cell types. But WARF argues in the suit that not only had the option expired a week earlier, but that it could also be denied at WARF's discretion.

Although any cell type can be derived from embryonic stem cells, some are more likely to yield commercially viable therapeutics than others. WARF is seeking to keep some of the more promising cell types available to other researchers.

"We couldn't figure out a way to resolve this without a third party," says WARF spokesman Andrew Cohn. He says WARF and Geron have been negotiating the option for seven months. **J. K.**

them into laboratories as quickly as possible. BresaGen and ESI also say researchers will be free to patent any invention they develop using the cells. The companies' profits will come from commercializing therapies that emerge from the research.

"First options have become the standard," says Mary Ellen Sheridan, associate vice-president for research at the University of Chicago. Researchers and their institutions rarely object because private companies are better able to handle the cost of clinical trials and marketing. And most options expire within a year, after which the researcher can take the invention elsewhere, Sheridan says.

WARE, by contrast, will try to recover its investment by exercising its broad patent rights to human embryonic stem cells. The foundation does not ask for the option to license researchers' inventions, because any company that markets such an invention in the United States will have to pay WARE for a commercial licence to the underlying technology.

But some intellectual-property experts say WARE's patent is so broad it may be challenged in court. "When you have a broad patent there is the possibility that broad claims will be held invalid and you will be left with a narrower patent," says Rebecca Eisenberg, a specialist in biotechnology patent law at the University of Michigan. ■

Riders on the storm provoke studies of Atlantic dust

Rex Dalton, San Diego

US researchers are planning an extensive study programme following reports that atmospheric dust may be transmitting microorganisms between continents.

Results collated at a workshop held on 14–15 August in Florida will be used as the basis for a grant application to the National Science Foundation early next year, participants say. They hope that the foundation will support the investigation as part of its bio-complexity programme.

Organized by the US Geological Survey (USGS), the workshop brought together 50 scientists from 18 institutions and government agencies.

Participants say recent studies suggest that clay dust from the Sahara desert — particularly from Mali — might be transporting fungi, viruses and chemical contaminants to the Caribbean, the Gulf of Mexico and the southern United States.

For instance, *Aspergillus sydowii*, which causes disease in Caribbean sea fans, is a possible culprit in the retreat of coral reef in the Caribbean. The fungus has been found in aerosols collected in the Virgin Islands,

which are hit by dust storms swept across the Atlantic Ocean from June to October.

Researchers are also studying dust as a cause of the rising number of asthma cases in Barbados and Puerto Rico. Eugene Shinn, a USGS geologist, has even suggested that long-range dust storms could transport the virus that causes foot-and-mouth disease in livestock. He admits there are no data to support this hypothesis, but says that satellite photos from February show a storm sending plumes of African dust over Britain shortly before the outbreak of the disease there.

Ginger Garrison, a marine ecologist at the USGS Center for Coastal Geology in St Petersburg, Florida, says research is planned to analyse dust samples in the mid-Atlantic during storm episodes to compare them with samples from Mali and the Virgin Islands. "We have found quite a load of microorganisms hitchhiking on dust particles across the Atlantic," she says.

The researchers also hope to coordinate with the Asian Pacific Regional Aerosol Characterization Experiment, which has studied dust movements from the Gobi Desert to the western United States. ■

Japan banks on tissue store for successful drugs

David Cyranoski, Tokyo

A Japanese foundation is set to open a major tissue bank by the end of this year, in a move that pharmaceutical companies hope will bolster drug discovery and clinical research in the country.

Researchers agree that Japan needs such a bank to provide ready access to useful tissue samples, and to bring it in line with what is available in the United States and Europe. But some worry that filling the bank will not be easy.

The bank will be run by the Japan Health Sciences Foundation, a government-affiliated organization based in Tokyo. Under the plan, tissues including liver, lung and kidney removed during surgery would be stored at a facility in Osaka built by the foundation last year. This would then supply tissue material to both pharmaceutical companies and academic researchers on a non-profit basis.

"The bank will greatly increase research productivity and help companies compete with those overseas," says Kazutaka Ichikawa, senior managing director of the Japan Pharmaceutical Manufacturers Association. Many drug companies already



Deposit account: researchers worry that Japan won't find enough donors for its tissue bank.

contribute money to the foundation, which has provided them with cell lines and DNA samples for several years.

An official at the health ministry, which oversees the foundation's activities, says the bank will initially depend on tissue collected from eight hospitals in the Tokyo area.

Researchers will use the tissues to experiment with candidate drugs that have proved effective on animals, but are not yet approved for use in human clinical trials, foundation officials say. "This is standard in the United States and Europe, but it is still immature in Japan," says Tetsuo Sato, chairman of the non-profit Human and

Animal Bridge Discussion Group.

But the tissue bank will only work if doctors can persuade their patients to provide the necessary, explicit consent. A 1996 law forbids the use of organs in research if they were taken from people who only gave consent for them to be used in transplantation.

As a result of the law, human tissue samples are in short supply in Japan, and researchers have to import samples from the United States or rely on overseas partners to do the work.

Ichikawa and others hope that the presence of an official bank and strict enforcement of informed consent guidelines will help to win the confidence of patients and doctors.

But Masanori Fukushima, an epidemiologist at Kyoto University, says that Japan's existing legal framework will make it difficult for the new bank to receive enough donations. "Guidelines on informed consent were only in place since last year and they have no binding power," he says, arguing that clearer laws are needed to punish doctors or researchers who violate patients' rights to privacy and informed consent. ■

German centre set to improve brain's image

Alison Abbott, Tübingen

The Max Planck Society (MPS) has pledged to support a major new facility for developing neural-imaging technology at Tübingen.

The centre, which could be up and running as early as 2003, will help to develop invasive and non-invasive functional imaging techniques. The MPS plans to invest up to DM60 million (US\$28 million) in the facility, which could become one of the world's leading centres of excellence in the fast-expanding field of brain imaging.

Brain-imaging techniques, including functional magnetic resonance imaging (fMRI), allow scientists to look at neurological function in living brains. But at the moment they do this indirectly, making use of magnetic effects to track the flow of oxygenated haemoglobin and so allowing researchers to deduce neural activity.

However, the Greek imaging specialist Nikos Logothetis, who will work at the new centre, has made considerable progress recently in developing ways to observe neural activity more accurately. Neuroscientists are optimistic that, with more powerful fMRI machines and further research into their use, they will rapidly advance the usefulness of the technique.

Head start

The decision by the MPS to fund the new centre so generously seems to be partly motivated by its desire to keep Logothetis in Germany. He was hired from Baylor College of Medicine in Houston, Texas, to become a director of the Max Planck Institute for Biological Cybernetics in Tübingen in 1996.

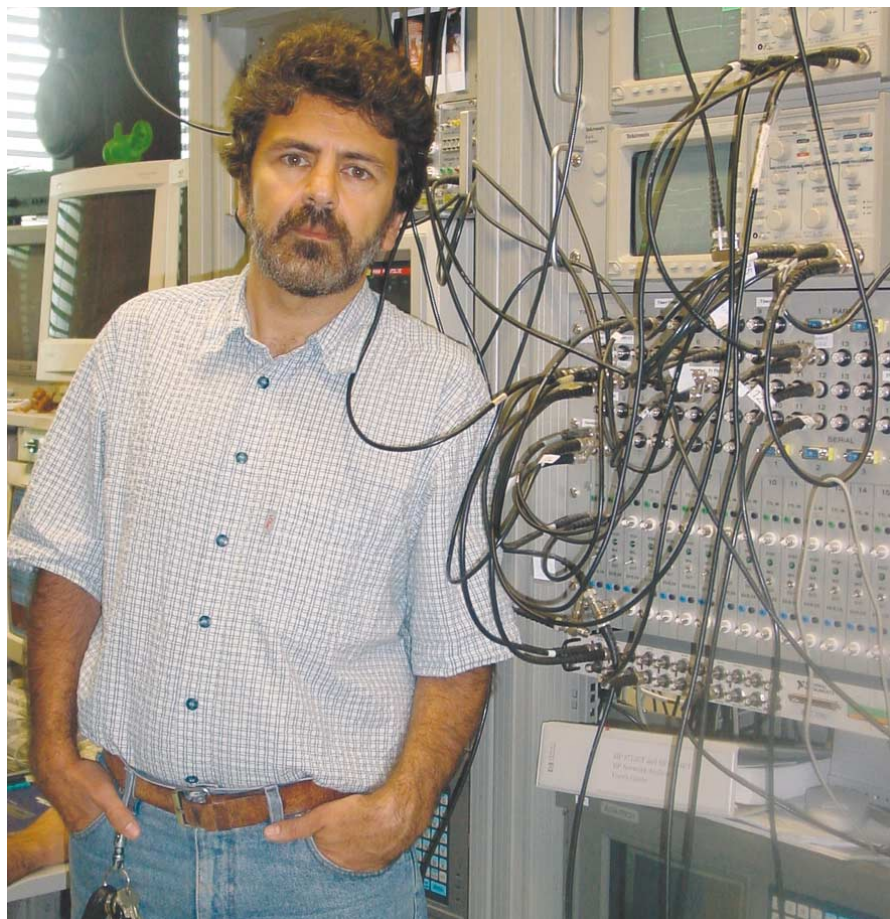
At the moment, the institute uses two magnetic-resonance machines, with maximum magnetic fields of 4.7 and 7.0 tesla, to analyse monkey brains.

As part of its work, Logothetis has helped to establish how the signal detected by the machines relates to neural activity.

Magnetic-resonance imaging — or nuclear magnetic resonance, as it used to be called — is widely used by doctors for anatomical imaging. It is also increasingly used experimentally for imaging the brain at work. But scientists are still learning how to interpret the images, which reflect the distribution of oxygen in the brain.

Researchers generally assume that high oxygen consumption reflects high levels of neural activity. But the new techniques developed by Logothetis use electrodes to monitor neural activity in monkey brains directly, while also using fMRI.

His most recent paper shows that fMRI signals correlate “unexpectedly well” with neuronal signals (see *Nature* **412**, 150–157;



Wired up: Nikos Logothetis plans to track the control and release of neural signals in the brain.

2001), but that measured signals routinely underestimate the extent of neuronal activity.

“The haemodynamic (vascular) signal is an important tool,” says Logothetis, “but the neuronal signal is more robust.”

He now wants to develop new technologies for neural imaging that measure the neural signal directly rather than through a surrogate.

Logothetis hopes to make use of newly emerging *in vivo* techniques for non-invasive monitoring of single molecules (see *Nature* **412**, 372–374; 2001). “These new techniques are just waiting to be applied to questions interesting for neuroscientists,” he says.

He plans, for example, to develop ways to track the calcium flow that triggers neural signals and to follow neurotransmitter release. This will require work by skilled physicists and chemists, as well as magnetic-resonance machines that generate strong magnetic fields. The MPS says it will provide these at the new centre.

The society is currently seeking a director with experience in this ‘high-field’ magnetic resonance to develop and run the centre. Logothetis has pledged to stay at Tübingen.

“The prospect of having access to such a centre on the campus where I work

was fundamental to my decision,” he says.

The centre will include a range of high-field fMRI machines, including very high-field machines for developmental work. A machine that generates a field of 15–16 tesla will be procured for work on small animals, and one of 7.0–9.5 tesla will be used for work involving humans.

Redrawing the map

The MPS hopes that Tübingen’s home state of Baden-Württemberg, as well as nearby universities, will help to extend the centre. The state government has already said that it is interested in participating.

“Proposing a 15-tesla machine is a very bold move,” says Robert Turner, a neuro-imaging expert at University College London’s Institute of Neurology. “The creation of the centre will change the map as far as imaging is concerned,” he adds.

Marcus Raichle, whose laboratory at Washington University in St Louis, Missouri, pioneered the use of positron-emission tomography for non-invasive imaging, says that Logothetis’s laboratory “has made enormous technical progress” in the past few years. Its expansion plan is “extremely exciting,” says Raichle. ■

Ecologists plot to turn the tide for shrinking lake

Quirin Schiermeier, Munich

The Aral Sea is dying. Researchers have given up hope that the central Asian lake — which was the fourth-largest in the world before intensive Soviet agriculture diverted the rivers that feed it — can ever be restored.

But a group of ecologists and economists is nonetheless discussing plans for a conservation strategy that could eventually reverse the ecological ruin of the region surrounding the shrinking lake. Germany and the United Nations Educational, Scientific and Cultural Organization (UNESCO) are backing the plan, for which scientific preparations could begin in Uzbekistan next year.

Over the past half-century, the Aral Sea has lost four-fifths of its volume and more than half of its surface area as water from the main rivers that flow into it, the Amu Darya and the Syr Darya, was diverted to irrigate cotton fields in the surrounding region.

The vanishing of the lake has triggered an ecological disaster, with desertification of what used to be the lake's floor and heavy salination of the surrounding water table and soil.

Intensive farming has also caused widespread pesticide contamination, according to scientists who have visited the region. Illnesses such as hepatitis, respiratory diseases and anaemia are widespread, and shortages of food and drinking water are worsening, a crisis that has been compounded by the exceptionally dry

summers of this year and last.

Now researchers at the Center for Development Research (ZEF) in Bonn are hoping to stop the rot by introducing sustainable patterns of land and water use to the region. They want to undertake a pilot scheme, after a preliminary investigation, to test the idea in the province of Khorezm, Uzbekistan, 400 kilometres south of the Aral Sea.

Under the scheme, around one-fifth of the irrigated cotton fields in the area would be converted into forest or hedgerow. More efficient farming would enable the remaining area to yield the same amount of cotton, the plan's architects say.

Christopher Martius, scientific coordinator of the project and an ecologist at the ZEF, thinks that local cotton growers can be persuaded to support the plan. "It is a legacy of Soviet times that most farmers in Uzbekistan have received solid scientific and agronomical training," he says. "Many of them have already signalled interest in participating."

The changes in land use would be accompanied by legal, administrative and economic reform, and its success will therefore require strong backing from the Uzbekistani government, which Martius also thinks will be forthcoming.

The initial phase of the project will prepare for the pilot by studying the ecology and economy of the region, including its hydrology, environment, demography and agriculture. This study will take up to four years, says Martius, after which a pilot will be implemented on a single farm and then, perhaps, throughout the 6,300 square kilometres of the province.

The new strategy aims to mitigate the consequences of the changing environment, rather than trying to rescue the Aral Sea. A UNESCO report released last year said that there was no prospect of preventing the sea from drying up.

"Changing the behaviour of people and authorities will not be easy," says Paul Vlek, one of the directors of the ZEF. "We hope, however, that we will be able to demonstrate that, even in a crisis situation, it is sensible to think ecologically."

Local scientists from Tashkent State Agricultural University and Urganch State University will be involved in the project from the outset. If, as expected, the German science ministry approves a grant of 3 million euros (US\$2.7 million) for the study phase, work on the project could begin next spring.

Code-breakers reveal results as threat of legal action is lifted



Rex Dalton, San Diego

Computer scientists who cancelled a talk in May because of legal threats from the recording industry delivered it in full last week at a conference in Washington.

The researchers, led by Edward Felten of Princeton University, presented their research on 15 August at the Usenix computer security conference, after the Recording Industry Association of America (RIAA) said that it did not intend to sue them.

The team had withdrawn virtually the same paper on code-breaking from an earlier conference, fearing legal action by the RIAA under the federal Digital Millennium Copyright Act of 1998, which prohibits certain disclosures of digital information (see *Nature* 411, 5; 2001).

In response to the case, a committee of faculty at Princeton will meet on 26 September to advise the university on possible policies to protect researchers' rights to publish, particularly in computer science. Some Princeton faculty have voiced concern that the university's leadership did not do enough to support the challenged researchers.

The new president of Princeton, Shirley Tilghman, was unavailable for comment. But an official at the university said that it strongly supports academic freedom, adding in a statement that university officials go "out of their way to indemnify our employees".

Members of Felten's team say that the legal threat has already impinged on their academic freedom. The group, backed by the Electronic Frontier Foundation, a California-based lobby group that supports free speech on the Internet, is now suing the RIAA in the federal court in New Jersey.



Washed up: ecologists can't save the Aral Sea but they hope to reduce the effects of its demise.

Staff suggest cure for Smithsonian woes

Josette Chen

Scientists at the National Museum of Natural History (NMNH) in Washington have come up their own plan for reorganizing research operations at the Smithsonian Institution, of which the museum is part.

The 13-page plan would divide science at the Smithsonian into three institutes of roughly equal size, dealing with astrophysics, natural sciences and environmental sciences. It would retain close ties between the administration of research and exhibits — ties the scientists have accused the Smithsonian's management of trying to break.

The 155-year-old institution has been in turmoil for most of this year as its secretary, Larry Small, clashed with its scientists over his plans to restructure research at the world's largest museum complex.

In May, the dispute led the Smithsonian's Board of Regents to appoint a senior scientific advisory panel, chaired by Jeremy Sabloff of the University of Pennsylvania Museum of Archaeology and Anthropology in Philadelphia (see *Nature* **411**, 624; 2001).

The scientists hope to influence Sabloff's panel when it meets for the first time next month. They propose setting up two Smithsonian-wide bodies: a directorate, comprising the directors of the three proposed institutes, to run research, and a council, including the heads of most research-performing branches, to advise on scientific objectives.

The Smithsonian Astrophysical Observatory at Harvard would make up most of



Only natural: researchers hope to retain links between science and exhibitions at the museum.

the astrophysics directorate, and the NMNH would dominate the natural sciences. The environmental-sciences directorate would combine several smaller centres.

Some scientists say the plan protects its authors' interests by giving the NMNH its own directorate. The authors respond that three directorates of about 200 scientists each would allow for efficient administration and benefit all researchers. The authors suggest appointing associate directors for research and for exhibitions at the major Smithsonian museums.

Brian Huber, an NMNH palaeobiologist and one of the authors, says that circulated

drafts of the plan were well received.

In response to criticism that their proposal does not go far enough, the authors say that: "Dismantling and reshuffling existing units is not necessarily a prescription for what ails Smithsonian science." They add that radical structural change could lead to years of painful upheaval.

David Umansky, the Smithsonian's director of communications, declined to comment. Sabloff says the proposal has been circulated to the panel members. "I welcome ideas and input from everyone," he says, adding that he would prefer to keep science and exhibits as integrated as possible. ■

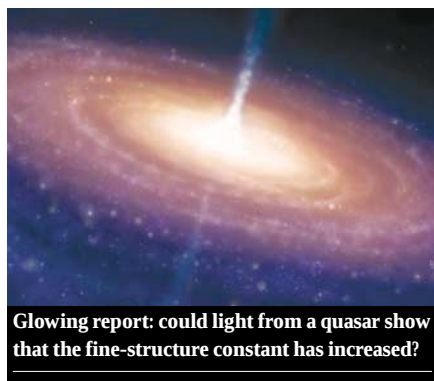
Flickering light raises possibility of changing 'constant'

David Adam, London

Physicists have reacted with curiosity and scepticism to suggestions that one of the fundamental constants of their discipline may not be so constant after all.

An international team, led by astrophysicist John Webb of the University of New South Wales in Sydney, made the claims after analysing light from bright, distant objects known as quasars. They say their results suggest that the 'fine-structure constant' — a measure of the strength of electromagnetic interactions — has increased by some 0.001% since the Big Bang.

If true, the finding would be revolutionary. Such constants determine the nature and interaction of matter throughout the Universe. If a shift in the fine-structure constant (also known as α) did take place, it would indicate that gravity and the weak and strong nuclear forces can also change over time. Such changes are impossible to explain using conventional theoretical physics and thus open the door



Glowing report: could light from a quasar show that the fine-structure constant has increased?

to rival explanations, including string theory, which invokes extra dimensions.

The result is statistically sound, but other physicists are concerned that a systematic error may be responsible for it. "One cannot see from the paper how they estimate the systematic uncertainties from various effects, only that they claim the errors are smaller than the effects they see," says John Bahcall, an astrophysicist at the Institute for

Advanced Study in Princeton, New Jersey.

Webb's team used data from the Keck telescope in Hawaii to study how clouds of interstellar gas absorb certain wavelengths of quasar light, producing a characteristic signature of dark lines in the spectrum that reaches Earth. They say the spaces between these lines in quasar light (effectively billions of years old) are different from those in similar spectra observed in the lab, and a variation in the fine-structure constant could offer an explanation.

"What we're claiming is that the sample we show is consistent with there being a smaller value of α in the past," says John Barrow, a physicist at the University of Cambridge. The team will publish its results on 27 August in *Physical Review Letters* (87, 091301; 2001).

Bahcall says that the complex analyses required will make it difficult for another group to test the claims. Barrow agrees: "This is not the sort of thing that people will be able to do in the next few months," he says. ■

Japan to train space laser on potential green fuel solution

Tokyo Plans to use an orbiting solar-powered laser to produce hydrogen for fuel cells are to be pursued by Japan's National Space Development Agency and the Institute of Laser Technology in Osaka.

The scheme proposes using a satellite to capture solar energy and power a laser trained on a solution of titanium dioxide and water back on the Earth. The laser would activate a process that separates the solution into oxygen and hydrogen. The hydrogen could then be used in fuel cells, a potential clean alternative to the use of petrol and diesel. Experimental simulations of the proposed laser system will begin this autumn.

Ultimately, researchers hope to use a satellite in an orbit at an altitude of 36,000 kilometres, high enough to track the Sun. The plan involves putting an experimental satellite in orbit by 2010, followed by a fully functional system in 2020.

Health balance tipped against landfill neighbours

London A study of the health of people living near landfill sites in the United Kingdom has shown a higher incidence of low birth weights and congenital defects among babies born to mothers living within 2 kilometres of the sites.

The survey found that rates of birth defects were around 1% higher than expected, a figure that rose to 7% near sites containing hazardous waste. The number of newborn babies weighing less than 2.5 kilograms was 5% higher than expected.

The team from Imperial College, London, that performed the study says that there are no known mechanisms that can explain the findings. The researchers adjusted their results to compensate for socio-demographic factors relating to populations around the landfills, many

of which are situated in areas of low income, but say that more work needs to be done before these factors can be fully discounted.

♦ <http://www.doh.gov.uk>

Boost for terascale computer research

San Diego In a move to realize the potential of supercomputing, the US Department of Energy last week announced plans to support 51 research projects in the area, at a cost of \$57 million for the first year alone.

The Scientific Discovery through Advanced Computing (SciDAC) initiative will develop software tools to allow researchers from a variety of disciplines to run simulations on terascale computers, which can carry out up to a trillion operations per second.

Of the new projects, 33 will focus on specific problems, such as climate modelling and fluid dynamics. The remainder will involve the development of software to improve researchers' access to supercomputers.

♦ <http://www.sc.doe.gov>

Sperm spinner succeeds in HIV-free fertilization

Tokyo Two Japanese women seem to be carrying healthy babies after undergoing *in vitro* fertilization (IVF) procedures using sperm from HIV-positive men, doctors announced last week.

The virus was removed before IVF using a method, developed by Japanese researchers, in which viral RNA or DNA within semen is broken down and separated from sperm by spinning it in a centrifuge (see *Nature* 408, 633; 2000). Healthy sperm are then selected and tested for infection. Blood tests show that the women, one due to give birth this autumn and the other next spring, are free of the virus.

The doctors say that this is the first confirmed creation of healthy embryos using sperm from HIV-positive men, although it is thought that previous

attempts have been made using a different technique. Researchers say the current method could be extended to remove other viruses such as hepatitis.

Fresh faces for British science funding

London New heads have been named at two of Britain's research councils, the country's main science-funding agencies.

Julia Goodfellow, currently head of crystallography at Birkbeck College, London, will become the first woman to lead the Biotechnology and Biological Sciences Research Council when she replaces Ray Baker early next year. "The UK research base in biotechnology and the life sciences is strong and internationally

competitive, but the rate of progress is so fast that we need to press vigorously over the next five years," Goodfellow says.

The Engineering and Physical Sciences Research Council last month announced that its new chief executive will be John O'Reilly, currently head of the electronic and electrical engineering department at University College London. O'Reilly will replace Richard Brook in October.



Julia Goodfellow and John O'Reilly.

Hum conundrum keeps scientists awake at night

Munich An investigation by German researchers has confirmed the presence of a nightly hum in several cities in the southern German state of Baden-Württemberg.

Groups of people living in Germany, Britain and the United States have reported being disturbed by low-frequency sounds during the past few years, but their complaints have often been blamed on hearing problems rather than external noise.

In response to complaints, Baden-Württemberg's environment ministry carried out nightly measurements in the homes of local residents. "When people told us the hum starts, we could clearly measure an acoustic signal," says Anja Bussek, a spokeswoman for the project.

The source of the drone — described as sounding like a distant engine or electrical motor — could not be determined. Furthermore, it remains unclear why only some people are able to hear the noise.



Laid waste: a study associates landfills with higher rates of congenital defects in babies born nearby.

The battlefields of Britain

With its farm-scale trials of genetically modified crops, Britain has taken ecological studies of farming practices into new territory. But the trials are the focus of intense controversy. Trisha Gura spoke to the scientists involved.



NIGEL CATTIN/HOLT STUDIOS

In the dock: oilseed rape is one of several crops that will be subjected to farm-scale trials in an attempt to evaluate the environmental impacts of its genetically modified counterparts.

Half a century ago, the British countryside echoed to the call of the skylark. Today, this much-loved bird is rarely heard. It was the fear that it and other farmland species might be silenced forever that in March 1997 led English Nature, a government conservation agency, to call for a moratorium on the introduction of crops genetically engineered to tolerate herbicides or to produce insecticidal proteins.

At a government conference in London to discuss developments in biotechnology, Brian Johnson, English Nature's biotechnology adviser, voiced concerns held by several groups. He blamed the long-term decline of bird populations, in part, on habitat destruction and intensive farming practices. He also

warned that genetically modified (GM) crops could make a bad situation worse. Crops resistant to broad-spectrum herbicides such as glyphosate were a particular concern, Johnson argued. If these relatively new agrochemicals denuded fields of their non-GM plant life, he said, they could remove the seeds and insects on which farmland birds feed.

The response of the incoming Labour government, elected in April 1997, was to delay commercial introduction of the crops pending the results of a four-year experiment costing £4.4 million (US\$6.4 million). Beginning with crop sowings in spring 2000, this unique study is comparing biodiversity in fields of herbicide-tolerant GM beet, maize and oilseed rape (canola) with that in comparable plots of equivalent non-GM varieties¹. Unprecedented in their size, these 'farm-scale' trials are designed to ask one basic question about one type of GM crop, explains Les Firbank of the Centre for Ecology and Hydrology at Merlewood in Cumbria, who coordinated the experiments. "The concern is simply: will the large-scale growing of these crops be damaging to wildlife?"

Firbank and his colleagues argue that the trials are the first attempt to investigate on an appropriate scale the ecological effects of an important change in farming methods in advance of its widespread introduction. The researchers hope to pave the way for similar

studies of other factors, such as pesticide use and tilling practices. "I see the farm-scale evaluations as being about developing techniques to assess the effect of proposed changes in land use and land management on biodiversity," says Firbank. "That makes it a very important study."

Trials on trial

But environmental groups and the organic farming movement view the trials in a very different light. They are implacably opposed to any GM plantings, commercial or experimental, and see the trials as an effort by the agribiotech industry to smooth the introduction of its new products. Mainstream environmental groups such as Friends of the Earth have attacked the trials' design as being scientifically inadequate, and more militant factions have attacked the experimental plots themselves.

The trials involve farmers growing one or more of the GM crops alongside its non-GM counterpart. Farmers who volunteered to take part sent their applications to a coordinating industry-backed body known as SCIMAC — the Supply Chain Initiative on Modified Agricultural Crops — which selected sites to be representative of a full range of British habitats.

For beet and maize, the trial will involve between 60 and 75 plantings of the GM crops,



Missing: skylark numbers have declined in Britain.



Harvesting data: Les Firbank hopes the farm trials will quantify the effects GM crops may have on plants such as the herb *Chenopodium album* (left).

each of about ten hectares in area, grown alongside matching plots of a non-GM variety. This breaks down to about 20 farms per crop, growing plots in successive years over the course of the study. "That should produce adequate statistical power to detect any differences," says Joe Perry of the Rothamsted Experimental Station in Harpenden, north of London, a statistician working on the trials.

For oilseed rape, there will be twice this number of plots—half sown as a winter crop in late August to mid-September, the remainder a spring crop, sown in March to May, depending on the weather conditions.

In each case, the non-GM crops are being treated according to the farmers' normal practices, including the application of conventional herbicides. The GM crops will be treated in the same way, except for herbicide sprayings. Here, participating farmers are being given a tighter protocol detailing the timings and quantities of sprayings with the broad-spectrum herbicides that the crops have been engineered to resist. This will be overseen by a scientific steering committee headed by Chris Pollock, research director of the Institute of Grassland and Environmental Research in Aberystwyth, Wales.

On the record

The GM and control plots are being monitored by a team of some 100 scientists, who are recording the species and number of key invertebrates such as slugs, butterflies and earthworms. They are also counting weeds and the presence of seeds in the soil. Although it was English Nature's concerns about birdlife that prompted the government to launch the experiment, only limited studies of avian biodiversity will be made. Birds range so widely, says Firbank, that individual fields are simply too small to study effects on bird populations.

But the researchers are confident that they can extrapolate the results on weed, seed and invertebrate biodiversity to bird populations. This will involve computer models such as those developed by Andrew Watkinson of the



University of East Anglia in Norwich. Last September, Watkinson's team reported that herbicide-tolerant sugar beet and the associated application of glyphosate could almost eradicate the wild herb *Chenopodium album*—fat hen or lamb's quarters—and severely diminish skylark populations².

Firbank questions these specific projections, noting that they depend on assumptions about parameters that will be measured in the farm-scale trials. But he believes that the ideas behind Watkinson's model could be developed and applied to the trial results to infer the knock-on effects on a variety of bird populations³. Johnson of English Nature

agrees: "We hope that the trials will produce data that will enable scientifically defensible decisions to be made about whether these crops will pose a threat to biodiversity."

But environmental groups question the decision not to monitor bird populations directly, and this is just one of a series of objections that they have raised.

First, they point out that the plots chosen were not monitored for biodiversity for several years before the trial began, giving no baseline data against which to track changes caused by GM farming. They also complain about the trials' specific focus, arguing that a chance to investigate practices in which little



Called to account: the trials will assess invertebrates such as earthworms (above) in a number of crops including sugar beet (top right) and maize.



or no herbicides are used has been missed.

"What we need is a proper agricultural research programme that looks at all of the options, including reduced dependency on chemicals," says Peter Riley, who heads Friends of the Earth's Real Food Campaign.

But the most loudly voiced complaints surround the possibility of herbicide genes being transferred to organic crops, or to weed species. The transfer of herbicide-tolerance genes to weeds is a long-running concern. Legal standards for organic produce — implemented by organizations such as the Soil Association, the biggest provider of accreditation to Britain's organic farmers — demand that it must be GM-free.

But scientists involved in the farm-scale trials argue that most of these objections have already been addressed. The lack of baseline data is not a serious issue, they argue, as the careful matching of experimental and control plots will ensure that meaningful comparisons can be made. Pollock and his colleagues add that they have looked at contamination issues, sponsoring a public forum to discuss the study design. Later, SCIMAC disqualified a farm near Coventry in the English midlands, some three kilometres from the headquarters of the Henry Doubtless Research Association, which conducts research into organic farming methods, fol-



Turf wars: Greenpeace activists are arrested for removing GM maize from a farm trial in Lyng, Norfolk.

lowing representations from the Soil Association and environment minister Michael Meacher. Two additional sites in Wales were abandoned after the farmers came under pressure from their local communities.

Over the years, investigations into gene transfer from GM crops to weed species have yielded mixed results⁴, depending on the crops involved and the presence or absence of closely related weeds. But a previous multi-year experiment, conducted at sites across Britain by a team led by Mick Crawley of Imperial College at Silwood Park, west of London, has shown that the transfer of genes to weeds from the crops involved in the farm-

scale trials should not pose serious problems. Crawley's team did not monitor gene transfer directly, but found no evidence that weeds were becoming more invasive, or surviving longer, at the experimental plots⁵⁻⁷. "There was no measurable difference in the ecology," says Crawley.

Design issues

Given the controversy surrounding the farm-scale trials, the scientists involved have taken the unusual step of submitting their study design to the *Journal of Applied Ecology* for peer review and publication. But this seems unlikely to placate groups such as Friends of the Earth, nor the more militant activists who have attempted to destroy several of the experimental plots.

The damage caused by these actions is difficult to quantify, says Firbank, because some attacks were limited to field edges, and for others the samples had been collected before the assault took place.

If the crop-destruction threatens to undermine the trials' statistical power, the study may be extended into subsequent growing seasons. But despite their determination not to be thrown off course by anti-GM activists, scientists working on the trials admit that the intensity of opposition has caused problems. "There is difficulty conducting the experiment within this hothouse atmosphere," says Perry.

But on one point, Perry and his colleagues can find some common ground with the more moderate of their environmentalist critics. Many of the scientists behind the trials would like to respond to Riley's call for a wider research programme investigating the ecological effects of other farming practices. Experiments into the effects on biodiversity of methods that rely heavily on pesticides, compared with methods featuring reduced chemical inputs, have been carried out at Rothamsted and other centres. But there has been nothing on the scale of the GM farm-scale trials. "Large-scale definitive studies cost a lot of money and there have been few, if any, of these," says Perry.

Pollock echoes Firbank's hope that the farm-scale trials will be the blueprint for

A green and pleasant land

Viewed from the United States, the surge of concern about farmland biodiversity that prompted Britain's farm-scale trials of herbicide-tolerant genetically modified (GM) crops is difficult to understand. Since the mid-1990s, US farmers have adopted GM varieties on an enormous scale, with barely a public murmur about the knock-on effects for wildlife — at least until concerns were raised about possible threats to the endangered Monarch butterfly (*Danaus plexippus*)⁹.

But the prairie-scale farms of the American Midwest have relatively little biodiversity to protect, except at their borders. And in a country that spans a continent, there is plenty of room to separate industrial-scale agriculture from areas of pristine wilderness.

"In the United States, we simply don't think in the same way because we have so much land," says Frank Forcella of the US Department of Agriculture's



research station in Morris, Minnesota, and the University of Minnesota in St Paul.

Although scientists such as Forcella are now planning studies of weed biodiversity in areas planted with herbicide-tolerant GM crops, the huge commercial plantings of these varieties in the United States have been accompanied by little in the way of studies into their effects on biodiversity. In the most part, the follow-up has concentrated on questions of yield and agrochemical use¹⁰.

Britain, by contrast, is a

crowded and heavily developed island in which the countryside consists of a patchwork of smaller fields, hedgerows, patches of woodland and human settlements. Agriculture, biodiversity, and the use of the countryside for leisure pursuits must all coexist. "We take much more concern with our farmland landscape much because we have so few truly wild places," says Les Firbank of the Centre for Ecology and Hydrology at Merlewood in Cumbria.

These factors explain why British ecologists have a long-standing reputation for studies of biodiversity in fields and hedgerows — which seem to be particularly important as havens for wildlife. Given this previous experience, the plots selected for the farm-scale GM trials include a range of different plots, from small fields surrounded by biologically rich hedgerows to the large arable fields of East Anglia that are Britain's nearest equivalent to the US 'cornbelt'.

future experiments investigating the influence on biodiversity of current and proposed farming practices. "This whole debate about the integration of agronomy into landscape is bubbling up in Europe right now," he says.

In fact, the design of the farm-scale trials could allow the investigation of one farming practice other than the use of herbicide-tolerant GM crops. Over the past three decades, farmers in northwest Europe have increasingly adopted a practice known as 'autumn sowing', in which crops are grown over the winter and harvested in spring. But ecologists are concerned that this practice, and its associated use of herbicides, has largely eliminated the 'weedy stubble' in which birds and insects find food and shelter in the early spring. As both winter and spring oilseed rape have been included in the farm-scale trials, it would be easy to adapt the trials to address this issue.

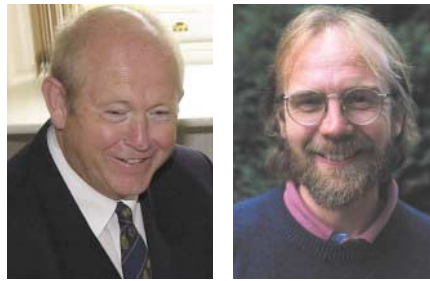
For agribiotech companies, such questions are of secondary interest to their hope that the trials will absolve herbicide-tolerant GM crops of the charge that they threaten biodiversity. GM proponents have long argued the reverse, pointing out that many farmers growing conventional crops use a range of different herbicides in startling quantities throughout the growing season — maize crops in Florida, for instance, can be sprayed up to 40 times.

Weeding out objections

In theory, the use of broad-spectrum herbicides in conjunction with GM crops engineered to tolerate their effects should mean that many fewer sprayings are needed. By using these crops, companies such as industry-leader Monsanto of St Louis, Missouri, argue that farmers can afford to wait until weeds are some 15 centimetres tall before spraying, providing enhanced habitat and food for invertebrates and birds. The herbicide-application protocols devised by Pollock's steering committee for the GM plots in the farm-scale trials reflect this recommended practice. Fewer sprays should also slow the emergence of herbicide resistance in weed species, say GM proponents.

The companies also argue that the GM technology will allow farmers to avoid tilling the soil at the beginning of every season — a practice that is meant to destroy weeds, but which is also thought to diminish biodiversity by reducing soil moisture and nutrients, and increasing the risk of erosion.

The British farm-scale trials are not designed to investigate tilling practices. But at Kansas State University, weed scientist Kassim al-Khatib is examining this issue in a much smaller study of soya beans and maize. This spring, he embarked on a four-year experiment, planting the crops at two sites of about 8 hectares each in Kansas, one dry and the other more humid. One-third of each site was planted with non-GM varieties, and sprayed with a conventional herbicide



Extended cover: Chris Pollock (left) hopes the trials can be developed to address other farming issues; a concern echoed by Peter Riley.

regime; the other two-thirds were sown with glyphosate-tolerant GM crops and sprayed with this herbicide when the weeds had grown to between 5 and 8 centimetres, or when they had reached 20 centimetres. Within each treatment, the plots were further split in half to investigate the effects of prior tilling versus no tillage.

But neither the farm-scale trials nor al-Khatib's experiment can predict what will happen once herbicide-tolerant GM crops get

technology since its introduction four years ago — 95% of the country's soya bean crop consists of herbicide-tolerant GM varieties. Scursoni fears that many farmers are spraying with herbicide more frequently than recommended. "Farmers tend to overcompensate," he says. "They see a clean field and they think it is a very nice result."

The scientists behind Britain's farm-scale trials accept that their study cannot answer all the questions surrounding the ecological effects of herbicide-tolerant GM crops — much less GM technology in general. But they hope that the principles of experimental design underpinning the trials will be used as a model for investigating these questions — plus a host of other issues surrounding farming practices — on a sufficiently expansive scale to yield conclusive results.

But the danger is that the controversy surrounding GM crops in Britain is now so intense that this message will be lost — and with it the chance to subject modern farming practices to rigorous ecological scrutiny. Even if the farm-scale trials fail to find GM



In opposition: the GM farm trials have met with significant resistance from environment groups.

into the hands of farmers away from the controlled conditions of a field experiment. Experience with maize engineered to produce an insecticidal protein does not bode well.

In May, John Obrycki at Iowa State University in Ames and John Losey of Cornell University in Ithaca, New York, reviewed a cluster of studies into the use of maize genetically engineered to carry a gene from the bacterium *Bacillus thuringiensis* that encodes a natural insecticide⁸. Obrycki and Losey concluded that most farmers growing the crops in the American Midwest were using the same amount of chemical insecticide as they had before the GM maize became available.

Julio Scursoni, a weed scientist at the University of Buenos Aires, and his colleague Eduardo Leguizamón at Argentina's National University of Rosario in Santa Fe, are now investigating whether farmers in Argentina are adopting a similarly aggressive approach with herbicide-tolerant crops. Argentinian farmers have enthusiastically embraced GM

crops guilty of harming biodiversity, suggests Crawley, it will make little difference to public perception. "Activists have so effectively demonized GM plants," he says, "that lay people now wouldn't believe a study that says there is no problem, no matter how well it has been done."

Trisha Gura is a science writer in Cleveland, Ohio.

1. Firbank, L. G. et al. *Nature* **399**, 727–728 (1999).
2. Watkinson, A. R., Freckleton, R. P., Robinson, R. A. & Sutherland, W. J. *Science* **289**, 1554–1557 (2000).
3. Firbank, L. G. & Forcella, F. *Science* **289**, 1481–1482 (2000).
4. Wolfenbarger, L. L. & Phifer, P. R. *Science* **290**, 2088–2093 (2000).
5. Crawley, M. J., Hails, R. S., Rees, M., Kohn, D. & Buxton, J. *Nature* **363**, 620–623 (1993).
6. Hails, R. S., Rees, M., Kohn, D. D. & Crawley, M. J. *Proc. R. Soc. Lond. B* **265**, 1–7 (1997).
7. Crawley, M. J., Brown S. L., Hails R. S., Kohn, D. D. & Rees, M. *Nature* **409**, 682–683 (2001).
8. Obrycki, J. J., Losey, J. E., Taylor, O. R. & Jesse, L. C. H. *BioScience* **51**, 353 (2001).
9. Losey, J. E., Rayer, L. S. & Carter, M. E. *Nature* **399**, 214 (1999).
10. Fernandez-Cornejo, J. & McBride, W. D. *Agricultural Economics Report No. 786*. (US Department of Agriculture, Washington, 2000); online at <http://www.ers.usda.gov/epubs/pdf/aer786>

► <http://www.defra.gov.uk/environment/fse/index.htm>

The positron probe

Beams of antimatter are providing some of the most detailed images yet of defects in semiconductors. Philip Ball reports.

An antimatter microscope sounds like an unlikely prospect. Microscopes that use beams of light or electrons to examine materials are well known devices. But matter and antimatter annihilate each other on contact, which makes an antimatter beam sound like a tool for destroying a sample, not studying it.

But in a paper published last month (*Phys. Rev. Lett.* **87**, 067402; 2001), Werner Triftshäuser and his colleagues at the University of the German Federal Armed Forces near Munich describe how they used a beam of antimatter particles to search for defects in a semiconductor. Their microscope, which uses positrons — the positively charged antiparticle equivalents to electrons — has attracted the attention of material scientists and electronic engineers because it can detect defects with a sensitivity of one part in a million.

At the atomic scale, most semiconductors consist of a repeating pattern of atoms. Defects are areas where this regular pattern is disrupted. Some types of defect can limit a semiconductor's ability to conduct electricity. If the number of these defects is too high, microchips made from the semiconductor will be useless.

Damage limitation

Around half of all the defects that damage semiconductors are vacancies — points in the lattice where an atom is missing. These vacancies can reduce the conductivity of a semiconductor by disrupting the flow of the electrons, and detecting them is a big challenge. Electron microscopes can provide nanometre-scale (10^{-9} metres) images of a surface, but they cannot detect vacancies. Other methods can identify vacancies at spatial resolutions of about a millimetre. Triftshäuser's device, although not able to spot individual vacancies, is the first probe to work at microscopic resolutions.

In the new microscope, a pulsed beam of positrons is trained on the sample being

studied. Positrons that enter the material are attracted towards electrons and repelled by atomic nuclei. The repulsion by nuclei means that many of the positrons diffuse into vacancies where nuclei are missing.

Ultimately, all of the positrons will encounter an electron and be annihilated, usually within a few hundred picoseconds (a picosecond is 10^{-12} seconds). But because most electrons in a semiconductor are localized around atomic nuclei, there are fewer

that the input pulses need to be just a few hundred picoseconds long. The positrons in Triftshäuser's microscope are derived from a radioactive sodium isotope and are then bunched into 200-ps pulses.

Sensitive scans

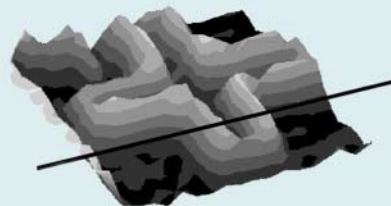
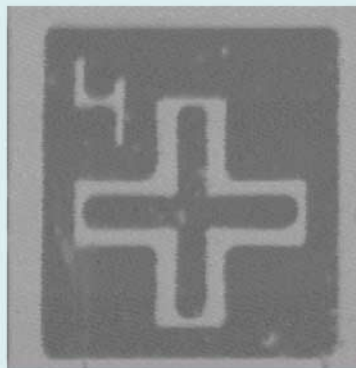
The Munich team demonstrated the microscope's sensitivity by detecting vacancies around a 75-micrometre-wide scratch in the surface of a gallium arsenide wafer, a commonly used semiconductor. They found that clusters of vacancies surrounded the scratch. Positrons survived for around 230 ps in defect-free areas, but lasted about 100 ps longer at the scratch's edge.

The microscope also has some capacity to differentiate between different elements. The team used it to scan a cross-shaped pattern of platinum, 120 micrometres wide, that lay on an oxidized silicon wafer. The new probe revealed previously unseen clusters of vacancies beyond the tip of the scratch.

The low intensity of the Munich team's positron beam means that a lot of pulses are needed to generate an image — it can take up to an hour to obtain a single pixel of the final picture. The researchers expect to improve on this by a factor of between 1,000 and 10,000 by using a more intense source from a nuclear reactor. The probe's spatial resolution, which depends on the diameter of the positron beam, will also be improved. The resolution is currently around 2 micrometres, but the team says that this could reach about 50 nanometres within the next five years.

The Munich researchers predict that their microscope will aid investigations into why manufacturing processes introduce defects into semiconductors and how these go on to damage chip performance. And because the accumulation of vacancies weakens metals by initiating cracks, the technique should find uses in measuring the strength and failure of these materials. The future looks bright for the positron probe. ■

Philip Ball is a consultant editor of *Nature*.



Revealed: platinum (dark areas) and oxidized silicon in a 120-micrometre cross imaged by the probe.



Antimatter architects: Werner Triftshäuser (left) and the team behind the positron probe (top).

electrons at vacancy sites than occupied sites. So a positron in a vacancy takes longer to meet and annihilate with an electron. This allows the researchers to use the lifetimes of the positrons to measure the number of vacancies — the longer the average lifetime, the higher the number of vacancies.

Lifetimes are measured by recording the delay between firing a pulse of positrons and detecting the gamma rays emitted as the result of positron-electron annihilations. But the short lifespan of the positrons means

How commercialization puts a blight on research

Good science stems from people with ideas, not from corporations or strategic teams.

Sir — The enthusiasm, discussed in your Opinion article “Towards a ‘knowledge nation’” (*Nature* **411**, 619; 2001), for commercialization of Australian basic scientific research, and the call for more entrepreneurial activity by scientists, needs tempering with a dash of reality.

Australian universities are in a parlous state, mainly because they have had little or no increase in real funding since the present right-wing government was elected in 1996. Lacking a tradition of private endowments, they are being encouraged to an unprecedented degree to seek commercial finance for projects.

Your Opinion article did not discuss many of the negative aspects of these changes. Commercialization of basic research will never lead to a ‘knowledge nation’, and basic research will not flourish in a commercial environment when profit dictates the direction of science.

The influenza drug Relenza was not invented by a pharmaceutical company, but resulted from years of curiosity-driven basic research in publicly funded institutions in Australia and overseas. These discoveries were then commercialized by GlaxoWellcome.

In our experience, commercialization of university research leads to diminution in the free flow of ideas, a focus on more applied projects and serious conflicts of interest. For example, past successes of the Australian National University (ANU) were used to float the company Biotron, set up to exploit discoveries, as yet unmade, in the medical sciences.

Of 65 million shares issued, 40 million are held by principals in the company, including scientists still employed by ANU but carrying out the research extolled in the company prospectus. This unacceptable conflict of interest appears to be tolerated, and indeed actively encouraged, by senior management in Australian universities.

To address the problems inherent in university research driven by commercial considerations, we need to identify the conditions required for good science to flourish. Good scientific research is not done by corporations, or by the strategic teams beloved of politicians and administrators, but through ideas which develop in the minds of individual scientists.

Strategic research will not produce truly novel discoveries. As George Porter, former president of the Royal Society, once said: “If a man comes to you with a strategic research plan, you’d better lock up the spoons”. US President Richard

Nixon had a famously comprehensive strategic plan to cure cancer within five years, half the time his predecessor, John F. Kennedy, had given NASA to land a man on the moon.

NASA achieved its goal, but 30 years after Nixon’s pledge we still have cancer. Why? Because we haven’t yet discovered what causes most cancers.

We will retain our best scientists and attract our most talented young people to take up a career in science only in an environment which encourages curiosity-driven research — and in order to flourish this needs to be free from the constraints of commercialization.

At the moment, Australian society is overwhelmed with examples of corporate greed and lack of ethics. The integrity of scientific research in universities must be

protected by staying at arm’s length from commercialization of its only product — the basic research that leads to a better understanding of the world.

Of course, scientific discoveries made in universities should be developed for commercial use and for the benefit of the country that funded the research — and scientists making those discoveries should be rewarded financially.

The ideal situation is vastly increased government support for curiosity-driven basic research and a mechanism, including an enforceable code of ethics, to commercialize any discoveries that are made in this way.

Graeme Laver, Arno Müllbacher, Paul Waring

Australian National University, Canberra, ACT 2601, Australia

Biotechnology gets big backing in Australia

Sir — The Opinion article “Towards a ‘knowledge nation’” (*Nature* **411**, 619; 2001) does not fully convey the priority that the Australian Commonwealth government — the national government of Australia — places on biotechnology.

For example, the Commonwealth government gave A\$296 million (US\$150 million) in the year 1999–2000 to biotechnology through funding schemes such as the National Health and Medical Research Council, the Australian Research Council and the R&D Start programme. This expenditure represents more than 9% of the government’s total R&D funding and is separate from the initiatives discussed in your editorial.

In limiting your article to individual initiatives, you did not discuss the coordinated nature of the Commonwealth response to biotechnology. This is best exemplified by the creation of Biotechnology Australia, a body that implements and evaluates national biotechnology strategy, and manages the government’s non-regulatory biotechnology activities.

The national biotechnology strategy was developed in consultation with the biotechnology sector and the government’s biotechnology consultative committee. It identifies priorities, including the Biotechnology Innovation Fund, set up with A\$40 million to address the market failure arising from the current shortage of

venture capital available at the critical proof-of-concept phase.

The government also has programmes such as the Innovation Investment Fund to encourage the flow of venture capital. The national biotechnology strategy also identifies the need for collaboration with state governments to facilitate the development and coordination nationally of existing clusters and networks. A national biotechnology centre of excellence is planned, with initial funding of A\$46 million.

Proponents of biotechnology must not forget that the public will not accept this new technology without appropriate safety measures and without information. The Australian Commonwealth office of the gene technology regulator started operations on 21 June, introducing stringent procedures to protect both human health and the environment, and serving the biotechnology sector by publishing clear guidelines and transparent procedures.

Commonwealth initiatives, such as labelling genetically modified food products and providing factual information for the public, continue to be a major activity (for example, see www.biotechnology.gov.au).

Sandy Radke

Biotechnology Australia, GPO Box 9839, ACT 2601, Australia

correspondence

contributions to correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, ideally shorter.

Galapagos or bust

A historical view of a busy scientific thoroughfare.

Evolution's Workshop: God and Science on the Galapagos Islands

by Edward Larson

Basic Books: 2001. 320 pp. \$27.50, £19.99

Joe Cain

Why are naturalists drawn so strongly to the Galapagos archipelago? Edward Larson claims that his "history of the Galapagos in science" will answer this question. Covering three centuries and hundreds of people, *Evolution's Workshop* is an ambitious book. Larson succeeds admirably with individual pieces of this story. He is a superb detective. But weaving these pieces into broader themes is rather more challenging. Although this is a brave attempt to study why scientists make use of a particular place, Larson's effort to build a bigger picture is clumsy and surprisingly timid.

With a straightforward chronology of expeditions to the Galapagos and related activities back home, Larson mixes the famous and forgotten, the professional and amateur. He describes interesting aspects of every visit studied by other historians and puts them into a meaningful sequence. Some visits were made simply to collect more data than a rival or to find data to resolve theoretical disputes. Others used the archipelago as a backdrop for testing new technologies, for new portrayals of nature or alternative political campaigns.

It wasn't Charles Darwin who made the Galapagos famous. As Larson makes clear, their fame is due to later naturalists, who shared the Victorians' excitement for exhaustive collection in foreign lands and



Eden under threat? Contact with humans has substantially perturbed the Galapagos biota.

their belief that working for science gave a nobility — a right to act — that trumped any other purpose. Crossing into the twentieth century, evolutionists saw simple ecosystems in Galapagos landscapes where variables could be isolated and theories tested. Then came conservationists, for whom saving the archipelago was a metaphor for saving the planet, and who constructed the image of an Eden under threat.

Larson has done splendid work in reconstructing the basics of who, when and why. He nicely splices the Galapagos expeditions

with the processing and theorizing back home that brought meaning to distant lands. Larson's coverage of late-nineteenth-century collecting and of conservation activities after the Second World War brings new information to Galapagos history.

But such a range brings inevitable quibbles. Most importantly, Larson emphasizes Darwin's zoological interests far too strongly when describing his visit. Rather, geology was centre-stage in the young man's mind. Besides, Darwin's brief visit was little more than a stop for provisions on the way to luscious Tahiti. He was later forced to borrow specimens and sift through sloppy notes when he subsequently thought something fundamental could be learnt from Galapagos mockingbirds and tortoise shells. Although Darwin later said the Galapagos was crucial to his thinking, this cannot be accepted at face value.

The splendid detail notwithstanding, Larson disappoints when he proposes larger meanings. He claims that the flow of Galapagos visits provides a lens for following "an epochal change in worldview" from subjective and allegorical to dispassionate and empirical. But his own account regularly contradicts these ideas. For example, the conservationists' recent allegory of Lonesome George, the last of a subspecies of Galapagos tortoises, was developed specifically to frame nature as sentimental. Larson also suggests that Galapagos history illuminates a transition between religion and science. But for this, he asks us to consider religion solely



No mates: Lonesome George (left) is the last of his kind and efforts to find him a suitable partner to preserve his genetic heritage have failed. Many species, such as this marine iguana, fared much better.





Not untouched: frigates and pelicans near a recent oil spill off the Galapagos archipelago.

in terms of a biblical literalism that spawned creationism, to ignore constant references to the Galapagos as Eden and Paradise, and to think of these two activities as polar opposites constantly at war. Such a framework is artificial and helps us understand nothing.

If Larson simply means that we should recognize how science influences the shaping of world views, he should have pressed further and considered its impact on notions of value. The original value of the Galapagos to humans was determined by its extractable commodities — fresh water, tortoise meat, seal fur and wood. Larson shows how value shifted when naturalists later came looking for things called 'specimens'. These were exported at extraordinary rates and in ever-widening variety. Exhausting the supply of specimens was avoided only because methodological fashions in biology changed, and specimens became worth more alive than dead. But there was nothing inevitable about this change. The large science budgets available after the Second World War, together with Western diplomacy, have paid generously to keep the Galapagos alive and intact as a biological preserve. Eco-tourism made preservation more profitable than anyone anticipated. Here, scientists have been both lucky and complicit.

An exploration of how science can influence policy on the Galapagos can illuminate similar negotiations elsewhere. Is this influence the result of science's growing cultural authority, as Larson claims? Or does it come from the money and political chips we're willing to trade for access? Galapagos history offers a rare opportunity to see scientists defend their values to policy-makers who might easily have acted to benefit other constituents.

Also notable is the persistence of an idea of the Galapagos that emphasizes pristine and unspoiled landscapes. This persistence is strange. Through four centuries of contact

with humans, the Galapagos biota has undergone sustained and substantial perturbations: native populations decimated, introduced species embedded and inter-island migrations enhanced. In what sense are these places untouched? This question has a bearing on our research preferences, too. Why do researchers value spaces "beyond the reach of humans" more than any other? What do we hope to learn from studying the pristine, given that human influences now are everywhere? Why do we seek Eden?

Evolution's Workshop offers a fascinating survey of science's contact with the Galapagos archipelago. Larson provokes many questions and provides rich historical detail. This book should launch important discussions about how science influences society and what we actually value when we choose to make nature precious. ■

Joe Cain is in the Department of Science and Technology Studies, University College London, Gower Street, London WC1E 6BT, UK.

Dispelling a myth

The Tangled Field: Barbara McClintock's Search for the Patterns of Genetic Control by Nathaniel C. Comfort
Harvard University Press: 2001. 352 pp.
\$37.50, £25.95

Mary-Lou Pardue

Barbara McClintock was a myth for me for years before I met her. In part, her mythical status was due to our mutual fascination with chromosomes. Her work on the break-age-fusion-bridge cycle and on the nucleolus



Legendary: Barbara McClintock was awarded a Nobel prize in 1983 for her work in genetics.

organizer were early influences on my interests in biology. Another part of McClintock's mythical status was her pre-eminence as a scientist and, of course, as a woman scientist. Hers was the first name mentioned by professors trying to assure me that women could have careers in science. (I never had any doubts that many women had the necessary talent, which made it harder to understand why I never saw women having such careers.)

My myth of McClintock was not the one with which Nathaniel Comfort introduces his account of her life — the myth of the obscure woman scientist whose work was disregarded by her peers for most of her life. But I agree that his is the more common, probably because most people are introduced to her before they are introduced to cytology and genetics rather than the other way round, as I was. After getting to know McClintock during the ten summers I taught a course at Cold Spring Harbor, I found my perception shifted somewhat, but my preconceptions retained much of their power — because they had a large factual component. Similarly, Comfort demolishes the common myth with authority, by taking us carefully through McClintock's life and work.

McClintock became a staff member of the Carnegie Institution Department of Genetics at Cold Spring Harbor, New York in 1942 and remained there until her death in 1992. Comfort did much of the research for this book at Cold Spring Harbor, itself a legendary centre for biologists. He also had access to McClintock's research notes, her correspondence, and her ears of corn (maize) with their multicoloured, streaked, spotted and sector kernels. He has talked to many people who knew McClintock — classical geneticists, young (and not so young) molecular biologists, people who worked at the laboratory and those who came to meetings there. The result is a multifaceted projection of professional views of McClintock's research, as those who were present remember it now.

Comfort asserts that "the social context cannot be understood without grounding it in the science" and the text proves his point. His description of the science begins with the birth of genetics. Today corn, with its long generation time and space requirements, gets passed over for more amenable model organisms — yeast, fruitflies, *Arabidopsis*. Comfort takes us through the early studies on corn, showing why it was one of the two dominant organisms used in pioneering studies of genetics. Studies on corn complemented the work being done with flies and raised questions that would fascinate McClintock for the rest of her life.

McClintock entered genetics early and impressively by discovering how to prepare corn chromosomes so that individual chromosomes and chromosome regions could be seen and identified. Her realization that

pachytene chromosomes could reveal morphological details not seen in more condensed chromosomes had practical value for mapping genes, but it also opened many questions about the meaning of these morphological features. (Of course, *Drosophila* polytene chromosomes present even more questions, but the way to visualize those chromosomes was yet to be discovered. If McClintock had worked on flies, might we have seen the wonderful polytene chromosomes years earlier?)

McClintock followed her corn chromosomes through their breakage and fusion antics, to a demonstration of the complexity of the nucleolus organizer, and then to the discovery of 'controlling elements'. Later, she used these chromosomes to study the origins and migrations of different races of corn. Comfort reconstructs the experiments, augmenting McClintock's difficult and minimalist published work with information from her field notes, memoranda and other papers. He presents a satisfying account of McClintock's discoveries of the dynamic nature of the genome. McClintock was always acutely aware that genetics needed to explain not only inheritance but also developmental biology and evolution. Comfort discusses her thinking on these subjects, pointing out that some of it was less than successful. Of course, even the greatest scientists sometimes bark up the wrong tree, but often just pointing out that the tree exists can influence later discoveries, even though that influence is not always recognized.

I heartily recommend this book to anyone interested in science or scientists. It will be welcomed even by those who have worked through McClintock's papers, because it fills in details omitted from her publications. For non-specialists, the science is clear enough to be accessible to anyone who has had a brief introduction to biology and genetics. The non-biologist has only to read around the experiments to become acquainted with a truly exceptional scientist. ■

Mary-Lou Pardue is in the Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, USA

Born in difficult times

Per Una Storia del Consiglio Nazionale delle Ricerche (2 vols)

edited by R. Simili & G. Paoloni

Editori Laterza: 2001

Vol. I 682 pp. IL70,000, 36.15 euros

Vol. II 688 pp. IL70,000, 36.15 euros

Giorgio G. C. Palumbo

The mathematician Vito Volterra, now best known for the solutions to integral equations that bear his name, played an impor-



Italian Nobel winners: left, Marconi (left), who wanted to transform the CNR into a great laboratory of war and peace, was admired by Mussolini (centre); right, Fermi's early work was supported by the CNR.

tant political role in Italian science during the early days of Mussolini's dictatorship. Respected at home and well-known abroad, Volterra did all he could to put Italian science on an equal footing with its international counterparts: he was instrumental in Italy's joining the principal international scientific unions (such as the International Astronomical Union and the International Union of Pure and Applied Physics).

So it was appropriate that, on 18 November 1923, Volterra became the founding father of the CNR, Italy's national research council. His dream was to create research institutes that could compete with the best of Western science. But Volterra was a Jew and was deposed as CNR president in 1930. After his refusal in November 1931 to swear loyalty to the Fascist government, he was forced to step down from all his academic responsibilities — as professor of theoretical physics at the University of Rome and as president of the Accademia dei Lincei. In December 1934, in a letter to his friend George Hale, then president of the US National Research Council, Volterra wrote: "If you no longer see my name in the list of members of the Accademia dei Lincei, do not assume I am dead." Sadly, the regime lasted longer than he did, and Volterra died in obscurity in 1940.

Such were the times in which the CNR was born. After almost 80 years of complex, often tormented history, the institution is now undergoing profound reform. So the publication of these two volumes (in Italian), in which 27 historians make a well-documented attempt to describe the intricacies of the CNR's evolution, is particularly well-timed.



The first volume covers the period from the CNR's foundation to the years of the Second World War (1939–45). Volterra's successor was the famous physicist Guglielmo Marconi, Nobel prizewinner and national hero, and much admired by Mussolini. Marconi promptly encouraged



Mathematical philately

A page of stamps featuring early Islamic mathematicians, from *Stamping through Mathematics* by Richard J. Wilson (Springer-Verlag, \$29.95), an annotated collection of mathematical stamps.

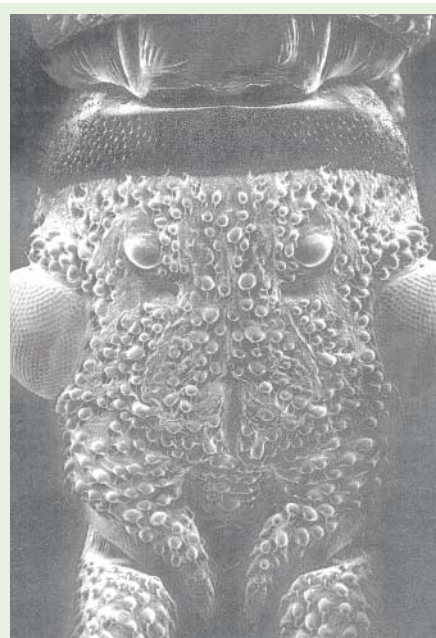
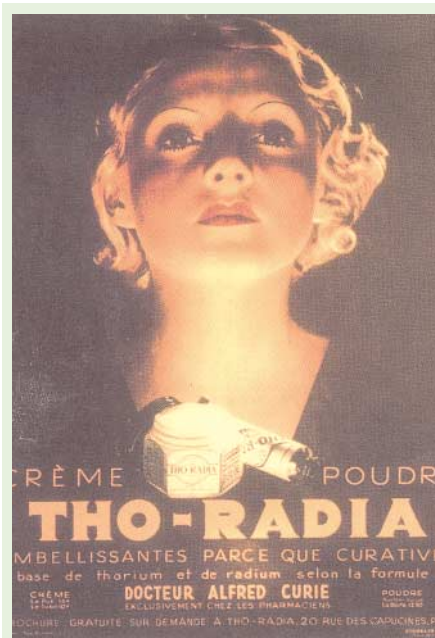
council members to transform the CNR into “a great laboratory of war and peace”. Soon after taking office, however, Marconi moved on to even more spectacular scientific adventures, leaving the CNR in the hands of the vice-president, a fascist bureaucrat. From 1937, and during the war, General Pietro Badoglio, a powerful military and political figure, became CNR president, with a mandate to promote war-orientated science and technology.

It is interesting that the Italian scientific community mostly ignored Badoglio's loud propaganda, and very little science for war was developed — unlike the situations in Britain and Germany. Mussolini did not trust university professors much, despising their inclination to pursue fundamental research and ignore industrial applications. With these unresolved contradictions, the CNR entered the war period.

This early story is documented in detail with fascinating extracts from correspondence between scientists, politicians and members of the CNR council. When read decades later, comments, opinions and gossip reveal much more about the political planning behind the scene than do the official facts — as well as revealing the personalities of the correspondents, all outstanding figures in Italian science.

For example, in 1936, Giulio Cesare Trabacchi, director of the Istituto Superiore di Sanità (National Institute of Health), lost 10,000 lire (a significant amount of money at the time) from his CNR grant after a clerk mistakenly assigned it to Enrico Fermi, who was then still working in Italy. Responding to a letter in which it was suggested that the sum could be returned to him and stripped from Fermi's grant the following year, he replied: “I have closely followed Fermi and his school's work and I well know how desperately meagre are the resources available to them. I would accept the proposition you suggest only if I could be sure my work could be compared to Fermi's and collaborators': I know this can't be done ... I therefore will continue with the ordinary grant I so far had access to.” It is now clear why Trabacchi is the only person Fermi thanked in his Nobel lecture after winning the prize for physics in 1938.

The second of these volumes spans the period from the post-war republic to the present. It is inevitably written almost as a chronicle, and touches on delicate points concerning people who are still alive and active in their fields. The change in style between the volumes prompts a general question: how will future historians assess our present era of non-archived web documents, e-mails and all-too-frequent mobile phone calls? Records of who was where at what time, and of what they were thinking or working on, and with whom — the



Beauty and the beast

Images from the *Dictionnaire culturel des sciences* (Editions du Regard/Editions du Seuil, 2001, FF495, in French), an eclectic mix of a thousand short articles on science and its influences on and from art, cinema, literature, politics and religion. From ADN (DNA to English speakers) to Zoroastre, this is a book to

delight and provoke, in which Niels Bohr keeps company with James Bond and William Blake. ‘Beauty’ (left) is an advertisement for a beauty product that purported to use radium; ‘beast’ (right) is an artist's impression of an electron micrograph of a bed-bug, entitled ‘Doctor Honoris Causa’.

human aspects of history — are all bound to be lost.

The post-war years have seen the CNR's scientists promoting European collaborations (through CERN — the European Organization for Nuclear Research — and ESRO, the latter of which metamorphosed into the European Space Agency and the European Southern Observatory, among other agencies), activating connections with the United States, particularly with colleagues who had fled the fascist regime, and starting technological projects, such as a programme for science in space. However, in the account, little emerges of the intricacies of the political scene. No doubt it is too early — the books, in fact, are presented as a contribution for a history of the CNR to be written in the future.

After the war, scientific leadership resided in the universities, and most CNR laboratory directors were university professors; CNR laboratories and research centres were physically located in university buildings. In the late 1960s, however, a significant change occurred with the creation of independent institutes; CNR directors are now mostly staff members of these facilities. The price of independence from academia has been high: scientific isolation and almost no access to graduate students and postdocs.

Since the 1960s, the CNR has had a peculiar policy of recruiting in ‘quantum’ jumps. A few months ago, for instance, about one thousand people became staff members all at once after about ten years of ‘hanging around’. But in spite of all the clumsy planning, a good deal of Italy's basic research is done within the CNR or in collaboration with CNR staff.

Present CNR leaders should take a close look at these two volumes. The political approach to science in Italy is described well, and its somewhat baroque design emerges clearly. All evolutionary efforts seem to have been in vain. For example, the most outstanding and constant feature of the CNR is its endemic lack of funds. The recurrent cry of presidents, of whatever political faith, has been that Italy's investment in science and research has been at best about half that of other developed countries, and more often only a third. As a rule, CNR scientists receive salaries, offices and lab space and not much else; research grants must be sought elsewhere. Perhaps the current restructuring will give the CNR the opportunity to finally realize Volterra's dream.

Giorgio G. C. Palumbo is in the Astronomy Department, University of Bologna, Via Ranzani 1, 40127 Bologna, Italy.

Yes, but what's it for?

The current state of language can make it difficult to discuss evolution in an accurate way.

Steve Blinkhorn

“These boots were made for walking,” sang Nancy Sinatra, and it's not too dangerous a metaphysical commitment to accept that that's just what they'll do. “These feet were made for walking,” though, is of an entirely different order, and is the sort of statement that brings palaeontologists and evolutionary biologists out in an uncomfortable rash. Evolution is not purposive — there is no teleology in phylogeny, only processes whereby the apparent order of living nature, although stable in the short term, is impermanent and without a certain future.

But wait. Henry Gee, in his account of cladistics, *Deep Time*, espousing precisely the view that we seek in vain for purpose in evolution, still writes: “In short, *Acanthostega* tells us that whatever limbs with digits evolved for, they did not evolve for walking on land” (p. 57). Is this Homer nodding, or is Gee just using a familiar form of language to draw out its own inapplicability? Notice the attribution of purpose in the last sentence. We accept purposive accounts of human behaviour in — perhaps especially in — the conduct of scientific controversy concerning the role or absence of purpose in evolution.

The truth is that language serves science very poorly in providing few clear, concise and cogent ways of indicating that randomness in a process does not necessarily lead to random results, so you do not always need purpose to account for order. We all know that, but communicating it is the very devil of a job. And, in English at least, it can take a good deal of effort to distinguish clearly between fitness for function and purpose intended, between what things are good for and what they were made for. Axes are good

Feathers are good for insulation, whether on their original owners or in a quilt, but no one supposes that feathers evolved to provide us with bedding.

for chopping wood, but was the first flint hand-axe made for the purpose, or was it just found to be fit for that use (if that's what hand axes were used for at all — itself a matter of some debate)?

Days after I wrote the first draft of this essay, a Letter to *Nature* described ‘feather homologues’ across the whole body surface of a fossil dinosaur that was clearly incapable of flight, raising more pointedly than ever the question: “What did feathers evolve for?” In point of fact, feathers are good for many different things, including flight, insulation (whether on their original owners or in a quilt, but no one supposes that feathers evolved to provide us with bedding), sexual display, writing, and burning beneath the noses of the putatively dead to check that they have indeed departed this life.

A little later, in a medical feature in a daily newspaper, a particularly slippery usage turned up: “Red blood cells have no DNA so they can transport more oxygen.” Intent, mere consequence, or what?

As every schoolboy used to have to learn, and some of us still remember, there is a very common construction in Latin involving a subordinate clause introduced by the conjunction ‘ut’ where the verb is in the subjunctive mood. “Crevunt pedes ut in terra ambulante animalia” translates both as “feet developed so that animals might walk on land” and as “feet developed, with the result that animals could walk on land”. It's as much hard work in Latin as in English to find everyday ways of describing processes leading to results without purposive intent being readily implied. And since Latin was the universal medium of scholarly communication in Europe for nearly two millennia, one may wonder just how much influence this one construction had.

Are there languages that have rigorously distinct ways of expressing purposive and non-purposive accounts of change processes? If so, do their speakers less readily accept purposive accounts in those areas where they are most tendentious? Are German speakers less prone to attribute purpose to evolution, for instance, or at least clearer as to whether they do or not, with the distinctive adjectives *bestimmt* and *geeignet* available to express the difference between ‘destined for’ and ‘fit for’?

There is a proposed link between the emergence of language and theory of mind, the propensity of people to act as though they believe others to share the same style of men-



Made for walking? Language often implies that the blind process of evolution has a purpose.

CORBIS

tal experience as themselves, and to attribute motives, intentions and expectations to them. One hypothesis concerning autism is that sufferers lack the ability to do this effectively or at all; autism is generally also marked by poor language skills.

If theory of mind and language really did develop in parallel in the course of human evolution, it would not be surprising to discover a bias towards explanatory forms deriving from mentalistic concepts such as intention, purpose and motive embedded so deep that it pervades all language and all languages. The kind of religious thinking that purports to discern divine intention in natural catastrophes can be seen as over-enthusiastic application of theory of mind to the physical world.

There is still some mileage in modelling molecules as billiard balls and atoms as planetary systems, however far cutting-edge physics has left these models behind. Modelling process as purpose may just be too deeply engrained in our forms of language for us to be able to relinquish it altogether. ■

Steve Blinkhorn is at Psychometric Research and Development Ltd, Brewmaster House, The Maltings, St Albans AL1 3HT, UK.

Delicate information

Rainer Blatt

Living in the 'information age', we are constantly concerned with a faster flow of information, made possible by the Internet, high-speed processors and communication links. 'Processor speed', 'memory capacity' and 'bandwidth' are terms that are known to us all, and our need for more of them drives one of the world's most flourishing industries.

Industry has long since described the rate of at which technology progresses in terms of a rule of thumb — commonly known as Moore's law — which simply states that processor speed, and hence computational power, doubles roughly every 18 months. This empirical law has proved to be valid over the past three decades and still holds today. It will not be long, therefore, before the number of atoms used to store a single bit of information becomes so small that the storage and processing of information will need to be described in terms of quantum mechanics. Eventually, a single atom will suffice for storing the information of a single bit, a state of affairs that will usher in the age of quantum computing.

With a single atom, information is stored in its quantum states, which can represent not only 0 and 1, but also otherwise rarely obtained 'superposition' states. In other words, the information may be found to be simultaneously 0 and 1, and a (quantum) measurement assigns it to either one of these states with a certain probability. This elementary type of information storage — termed a quantum bit or 'qubit' — seems to be an odd proposition, because at first it seems impossible to handle and process such elusive quantum information in a predictable way. Therefore, in the 1980s and the early 1990s, qubits and quantum information processing were considered to be a mere curiosity in the information sciences.

This point of view was shattered when Peter Shor of AT&T came up with a stunning algorithm in 1994. He showed that the factorization of large numbers — an extremely difficult and time-consuming problem for

classical computers — can be achieved in a highly efficient, rapid way using a quantum computer. Large numbers and their prime factors are commonly used to encrypt data and messages, as it takes too long to factorize a large number — and thus decode the information — on a classical computer. If the factorization problem could be solved, then all current cryptographic systems would be seriously endangered. Peter Shor's result therefore spawned a worldwide attempt to create a real quantum computer, which could be applied to basic research and technology.

After more than five years of research, the field of quantum information is now thoroughly established, and earnest investigations are being made of about ten different technologies that may become the computing techniques of tomorrow. Is this the beginning of a new computer age — the birth of a new technology for number-crunching? Or is quantum information something more profound — perhaps a completely new concept — that is intellectually worth increased research efforts?

Although quantum computation is often hailed as the computing technology for the twenty-first century, it has become clear in recent years that today's computers will not be replaced by quantum counterparts. Aside from factorization, only a few algorithms run faster on a quantum computer, and the classical workhorses of information processing are set to remain in place for the foreseeable future. That favourite question of journalists — "when can I buy a quantum computer?" — is definitely beside the point here.

At best, current quantum technology can provide a system with a few (less than ten) qubits and a limited number of the processing operations that are required for computation. But progress over the past few years offers real hope that we will eventually be able to control and manage the technology to build and operate larger-scale quantum computers with a high degree of reliability. Comparatively speaking, however, at present we are not yet even at the stage of the ENIAC (electronic numerical integrator and computer), the 1940s predecessor of

Quantum computing

Is this the birth of a new technology for number-crunching, or is quantum information something more profound?

modern computers.

Nevertheless, viewing the fundamental physical laws of quantum mechanics from the angle of quantum information has had a remarkable impact so far. Five years ago, the ingredients of quantum computing technology were seriously thought to be uncontrollable. Handling and processing quantum information — that is, controlling the superposition states for extended periods of time and over ever-increasing distances — still requires very delicate procedures. Every unwanted interaction of a quantum storage device, such as an atom with its surrounding environment, would constitute a measurement and would thus lead to irretrievable loss of that information.

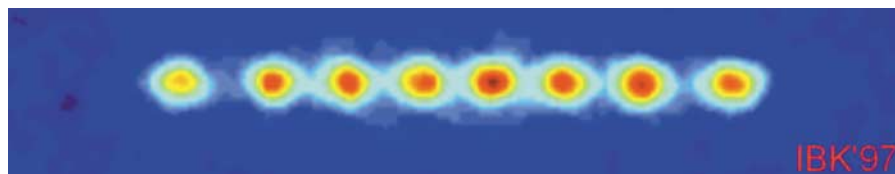
Nevertheless, research has shown that correction of errors at the quantum level is a genuine possibility, and can be used to protect and even to restore quantum information. This is a surprising and revolutionary result that was not foreseen even as recently as a few years ago. Aside from its potential technical impact, it clearly demonstrates the power of viewing fundamental physics with the information-guided eye. Quantum computation, and more generally, quantum methods for information processing and communication, not only provide new tools for the information age, but also represent one of the most powerful concepts since the discovery of quantum mechanics itself.

The schemes and techniques being investigated today will be the fundamental building blocks for tomorrow's general quantum technology. Thus, the control and measurement of large-scale quantum systems — what will come to be termed 'quantum engineering' — may be the key technology of the twenty-first century. ■

Rainer Blatt is at the Institute for Experimental Physics, University of Innsbruck, Technikerstrasse 25, A-6020 Innsbruck, Austria.

FURTHER READING

Bouwmeester, D., Ekert, A. & Zeilinger, A. (eds) *The Physics of Quantum Information* (Springer, Berlin, 2000).
Steane, A. *Rep. Prog. Phys.* **61**, 117–173 (1998).
Bennett, C. H. & DiVincenzo, D. P. *Nature* **404**, 247–275 (2000).



Counting qubits: a string of ions representing a 'quantum byte' in an ion-trap quantum computer. Here all are in state '0'; exciting any one would cause loss of fluorescence at the corresponding position.

Fast track to fusion energy

Michael H. Key

Nuclear fusion could solve the world's energy problems, but its potential remains untapped. Can a new way to ignite a tiny ball of high-density fuel make the dream come true?

California's recent energy shortage has highlighted the potential fragility of the energy supply in a world of increasing economic activity. The reservoir of fossil fuels will eventually be exhausted, but even if it were limitless, global warming from CO₂ emissions makes reliance on fossil fuels perilous. Energy sources with no CO₂ emission are therefore important. One of the most attractive of these is nuclear fusion, but obtaining energy from this process has proved difficult. On page 798 of this issue Kodama *et al.*¹ report the first demonstration of a new concept in fusion called fast ignition, which may offer a better path to fusion energy.

Nuclear fusion powers our Sun, the stars and thermonuclear weapons, so what's stopping it being used as an energy source? The answer has long been 'nothing in principle'. The deuterium isotope of hydrogen (²H) occurs naturally in water, providing an essentially unlimited fuel supply. Its fusion with the tritium isotope, ³H, has the greatest yield of energy per mass of fuel (17.6 million electron volts per fusion event) of any energy source under consideration. Although tritium does not occur naturally, it is readily produced from naturally abundant lithium in a breeder reaction induced by neutrons from the fusion process. Moreover, in contrast to nuclear fission, the D-T fusion fuel itself has no long-lived radioactive decay products. One problem for fusion energy is that it makes the reactor structure radioactive, but this can be mitigated by using construction materials of low atomic number.

The main obstacle to progress is that thermonuclear fusion requires a temperature of 50 million °C, which presents scientific and technical problems to producing fusion under controlled conditions. Two avenues are being explored. In magnetically confined fusion energy (MFE), hundreds of cubic metres of D-T plasma at a density less than 10⁻⁹ g cm⁻³ are confined by a magnetic field at a pressure of less than 10 atmospheres and heated to the fusion temperature. The aim is to create a steady-state burning plasma to which fresh fuel is added, as in a furnace.

In inertial confinement fusion (ICF), on the other hand, a spherical shell of solid D-T ice a few millimetres in size is compressed in about 10⁻⁸ s to hundreds of g cm⁻³ at a pressure of more than 100 giga-atmospheres.

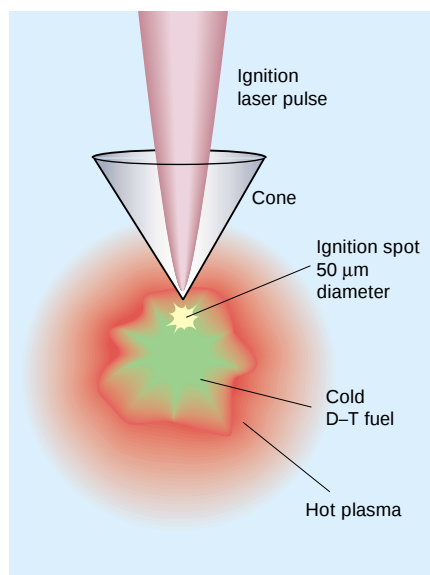


Figure 1 The fast-ignition concept. A roughly spherical mass of compressed deuterium-tritium (D-T) plasma is formed by the implosion of a hollow spherical shell of D-T around a hollow metal cone. A petawatt laser focused at the apex of the cone creates an intense beam of electrons moving at almost light speed, which ignites the D-T fuel. Kodama *et al.*¹ tested this idea for the first time in a scaled-down sub-ignition experiment using a deuterated polymer to simulate the D-T fuel pellet. Their experiment uses a short 60-joule laser pulse (0.06 petawatt, 1 picosecond duration) to increase the number of fusion events tenfold relative to the same experiment without the 60-joule pulse.

Ignition is produced in a small, central hot spot and thermonuclear burning consumes the rest of the fuel, the reaction being confined for less than 10⁻¹⁰ s by inertia alone. The aim is to produce repeated microexplosions, as in an internal combustion engine.

MFE has been researched for more than four decades. So far its twin pinnacles of achievement are the demonstration of 11 megawatts of fusion power generation for 0.7 s and 16 megawatts for 1 s by facilities in the United States and Europe. In both these tests the fusion power output was significantly less than the power input, but these and other studies have been used to design machines that would reach burning-plasma conditions with a net excess of fusion power over input power. International negotiations

on how to evaluate the various designs and share the cost of such a project are currently in progress.

By comparison, ICF is still a young field. It uses a driver — either a high-powered laser or ion beam — to compress and ignite the fuel. Currently under construction in the United States and France are stadium-sized lasers, which will generate laser pulses with energies of up to 2 megajoules in a few nanoseconds. They are designed² for ICF with a ratio of fusion energy to laser energy of about 15:1. Fast ignition³ is a new variant of ICF that could both produce a much higher ratio of fusion energy output to driver energy input and reduce the precision required for the driver and fuel pellet. Although fast ignition is still at an exploratory stage, these features make it attractive for building ICF power plants⁴.

The current megajoule laser projects require exceptional uniformity of the drive pressure, and extremely smooth and spherical fuel capsules. Such high precision is needed to produce the central ignition hot spot. The necessary conditions can, in principle, be achieved in single-event experiments but would present technical and economic problems for routine operation of an ICF power station with ignition events occurring ten times per second.

The fast-ignition concept avoids these difficulties by separating the processes of fuel compression and hot-spot generation (Fig. 1). A laser or ion-beam driver first compresses a spherical shell of fuel. Then a 10-petawatt (10¹⁶ W) laser⁵ pulse of 10 picosecond (10⁻⁸ s) duration is focused on the fuel to create a hot spot at its edge at the instant of peak fuel compression. The heating is caused by laser-accelerated electrons moving at near the speed of light, or, in a more recent proposal⁶, by laser-generated protons. Fast ignition uses fuel densities five times lower than conventional ICF. The lower energy invested in compression allows a higher ratio of fusion energy output to driver energy input of about 300:1. The reduced fuel pressure and absence of the central hot spot also mean less stringent requirements on the smoothness and sphericity of the fuel capsule and uniformity of the drive pressure. But our understanding of the physics involved in the fast-ignition concept is far less mature than that for conventional ICF. More work on relativistic

laser-plasma interactions and propagation of giga-ampere currents of megavolt electrons is required⁷. The efficiency of energy transfer from the short laser pulse to the hot spot in the compressed plasma is a crucial parameter, which has not yet been accurately measured or predicted, but must reach about 20% for the scheme to succeed.

The experiment by Kodama and colleagues is the first to combine production of compressed matter in a laser-driven implosion with picosecond heating by a laser pulse timed to coincide with peak compression. This is a small-scale test of the fast-ignition concept and uses a small spherical shell of deuterated polystyrene to simulate the D–T ice. Because the target has no tritium, it produces D–D fusion instead of D–T fusion. The shell is formed around a hollow gold cone to provide an access path for the ignitor beam to the fuel, which is compressed to a diameter of 40 micrometres at 50 g cm^{-3} (Fig. 1).

Their first results are encouraging. The thermonuclear yield of neutrons from D–D fusion is increased tenfold by irradiation at peak compression with a short 60-joule laser pulse. This means that 1% of the temperature rise needed for ignition was achieved using only 0.1% of the theoretically required ignitor beam energy. They also note that the longer, 1,200-joule laser pulse that drives fuel compression has to be increased to 2,800 joule to achieve the same increase in neutron yield without the 60-joule short-pulse heating — highlighting the efficiency of the fast-ignition approach. Preliminary analysis suggests greater than 20% energy transfer efficiency from the short pulse laser to heating of the compressed core, which is adequate for full-scale fast ignition.

If this heating efficiency were confirmed by experiments using 1-petawatt, 500-joule ignition pulses, planned by Kodama and co-workers for later this year, the longer-term prospects for full-scale fast ignition would be good. Stepping up to megajoule laser drivers, it might even lead to a 300-fold energy gain, and could initiate serious efforts worldwide to produce fusion energy by fast ignition. However, experience with other fusion energy projects should temper our enthusiasm for fast ignition. At such an early stage, this new approach to fusion energy should be viewed as promising, but speculative, until much more work has been done.

Michael H. Key is in the Lawrence Livermore National Laboratory, University of California, PO Box 808, Livermore, California 94550, USA. e-mail: key1@llnl.gov

- Kodama, R. *et al.* *Nature* **412**, 798–802 (2001).
- Lindl, J. D. *Phys. Plasmas* **2**, 3933–4024 (1995).
- Tabak, M. *et al.* *Phys. Plasmas* **1**, 1626–1634 (1994).
- Hogan, W. J. (ed.) *Energy from Inertial Fusion* (IAEA, Vienna, 1995).
- Perry, M. D. & Mourou, G. *Science* **264**, 917–924 (1994).
- Roth, M. *et al.* *Phys. Rev. Lett.* **86**, 436–439 (2001).
- Honda, M., Meyer-ter-Vehn, J., Pukhov, A. *Phys. Rev. Lett.* **85**, 2128–2131 (2000).

Neurobiology

The many faces of adaptation

Pamela Reinagel

Sensory neurons constantly adapt to the changing environment. It seems they can do this surprisingly quickly, and without losing their absolute frame of reference.

Sensory nerve cells can produce only a small number of distinct outputs, so they cannot generate a unique response to every relevant stimulus. To solve this problem, neurons adapt to the prevailing conditions, so the same limited set of outputs can be reassigned to different stimuli in different contexts. It was already known that, after a sudden change in the strength of a stimulus, neurons abruptly change their average responses (firing rate). Then, as the cells adapt to the new conditions, the mean response gradually adapts back to a more moderate level. Fairhall and colleagues¹ (page 787 of this issue) now show that this change in average response is just part of the story: different aspects of a neuron's response adapt independently on different timescales. Uncoupling these aspects provides surprising insight into the functions of adaptation.

When you walk from bright daylight into a dimmed auditorium, you are initially almost blind, but after several minutes your eyes adjust and you can see quite clearly. Emerge from the auditorium an hour later and the sunlight is momentarily blinding. This familiar experience illustrates the importance of sensory adaptation: without it, we would be effectively blind except at one specific level of illumination. Instead, vision can operate over a 10^{11} -fold range of light levels (from about 10^{-6} candela per square metre in darkness to 10^5 in full sunlight), even though only 10 to 100 distinct values can be distinguished in any given context.

The level of light is just one of many features to which sensory systems adapt. For example, nerve cells that respond to a more complex feature, such as motion, can also adapt, in this case to the amount of motion in a scene. Neurons adapt not only to the average but also to the distribution (variance) of recent stimuli. Both kinds of adaptation help the cell to discriminate between typical stimuli in a given context (Fig. 1).

It can take several seconds for the average firing rate of a cell to adapt after a change in the variance of a stimulus. Is this just a result of the inefficiency of biological mechanisms? Perhaps not. Cells adapt more quickly when variance increases than when it decreases². This might be explained by a problem inherent to measuring distributions — how many stimuli would you need to observe before you were convinced that they came from a new distribution? Consider the situation

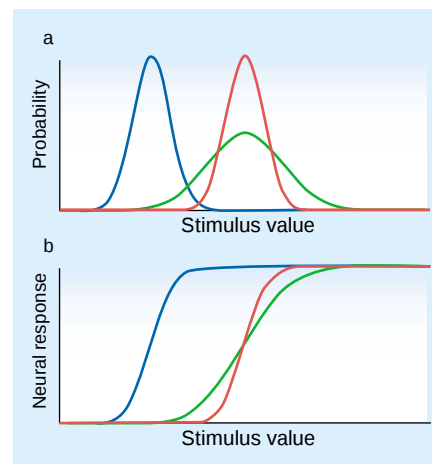


Figure 1 Natural stimuli change dramatically in different contexts; for example, light levels vary from day to night. Sensory neurons adapt to such changes. When average stimulus values (such as light intensity) are low (a, blue curve), a neuron uses its limited range of responses to represent the most likely values (b, blue curve), leaving it unable to distinguish among lower or higher values. If the average stimulus value suddenly increases (a, red), the neuron adapts by shifting its response function to the right (b, red). This optimizes encoding of the new stimuli. If the average value is held constant but the range suddenly becomes broader (a, green), the neuron becomes less sensitive to small differences (b, decreased slope of green curve), so the whole range of likely values can be represented. The neural response is shown as a continuous variable such as firing rate, but the same principles hold for more complex properties such as firing patterns. Fairhall *et al.*¹ reveal that adaptation involves many processes, occurring at different rates and with distinct functions.

in Fig. 1a. When variance has been low (red curve), just a few extreme values (from the green distribution) would quickly alert you to a change in the stimulus context. But when variance has been high (green curve), it would not be surprising to observe a few stimulus values from the red distribution; it takes a while before the lack of more extreme values becomes suspicious³. This raises the question of whether the speed of the change in firing rate (or of any other form of adaptation) is limited only by the time needed to measure the new distribution of signals.

Fairhall *et al.*¹ have now tested this by investigating how a motion-sensitive neuron

in the fly *Calliphora vicina* adapts its response when the variance of motion signals suddenly changes. Previous studies showed that after neurons have adapted to a change in variance, information transmission — that is, the ability to discriminate among stimuli — is optimized to the new conditions^{4,5}. Fairhall *et al.* find that the neuron adapts its input–output relationship extremely rapidly, achieving optimal information transfer within tens of milliseconds. The authors demonstrate that this is about as soon as any system could reliably know that the stimulus distribution has indeed changed.

On the other hand, as mentioned above, firing rate adapts much more slowly, over several seconds. So this cannot be the main mechanism by which information transfer is optimized. Instead, the authors show that the speed of firing-rate adaptation scales to match the frequency of context changes. That is, the dynamics of firing-rate adaptation depend not only on the variance of the preceding and following stimuli, as described above, but also on the ‘metastatic’ of how frequently the stimulus conditions change. This suggests that the brain could use the speed of firing-rate adaptation to predict how soon conditions are likely to change again.

Adaptation has a potential drawback, however: if a sensory neuron changes its input–output relationship without notifying the brain that it is operating in a new context, its messages will become ambiguous. In some systems this might not matter, because information about the context would be biologically useless⁶. But Fairhall *et al.* find that, in their system, information about context is not lost as a result of adap-

tation. Some firing statistics adapt to carry reliable information about the context (variance). These signals could be used to indicate which input–output relationship is operating, thereby removing any ambiguity arising from the shifting response properties of the neuron.

Fairhall *et al.*'s study¹ shows that many independent adaptation mechanisms occur, over a continuous range of timescales. Of course, this raises questions about the mechanisms underlying sensory adaptation. Different forms of adaptation could theoretically be implemented at different levels of organization, ranging from molecules to cells to neural networks. But the authors provocatively speculate that all the observed timescales might be mapped to different biophysical properties of sodium channels. Finally, it is often argued that the coding properties of sensory neurons are adapted to represent their natural stimuli efficiently. Such arguments need to consider that this adaptation may not be hard-wired but rather a dynamic process that occurs throughout the life of an organism. ■

Pamela Reinagel is in the Department of Neurobiology, Harvard Medical School, 220 Longwood Avenue, Boston, Massachusetts 02115, USA.

e-mail: pam_reinagel@hms.harvard.edu

1. Fairhall, A. L., Lewen, G. D., Bialek, W. & de Ruyter van Steveninck, R. *Nature* **412**, 787–792 (2001).
2. Smirnakis, S., Berry, M. J. II, Warland, D., Bialek, W. & Meister, M. *Nature* **386**, 67–73 (1997).
3. DeWeese, M. & Zador, A. *Neural Comp.* **10**, 1179–1202 (1998).
4. Brenner, N., Bialek, W. & de Ruyter van Steveninck, R. *Neuron* **26**, 695–702 (2000).
5. Wainwright, M. *Vision Res.* **39**, 3960–3974 (1999).
6. Clague, H., Theunissen, F. & Miller, J. P. J. *Neurophysiol.* **77**, 207–220 (1997).

Earth science

Journey beneath southern Africa

Suzanne Y. O'Reilly

Seismic analyses of the lithosphere, undertaken as part of the Kaapvaal project, provide an unprecedented view of cratons — the earliest parts of continental landmasses.

How continents formed on the early Earth is one of the big questions in geoscience. One way to tackle it is by studying cratons. These are the nuclei of continents that formed during Archaean times, at least 2.5 billion years ago. Southern Africa is an especially rich area of study, so it was the setting for the Kaapvaal project. This international and multidisciplinary programme was designed to probe the architecture and age of the region's lithosphere (the upper mantle and crust that constitute the upper 200–300 km of the Earth). The result, as described in papers in *Geophysical Research Letters*^{1–7}, is the most detailed picture yet of cratons.

Southern Africa is made up of several terrains ranging in age from Archaean to very young. One of the main sources of information about what is happening deep in the lithosphere is xenoliths — fragments of mantle rock that were carried to the surface in molten rocks called kimberlite magmas. Analyses of xenoliths that erupted through the lithosphere from 1.2 billion to 80 million years ago have allowed measurement of the ages (using a system based on the ratio of rhenium and osmium isotopes)^{1,2} and physical properties³ of the deep Earth beneath southern Africa at the times of eruption.

The most remarkable results of the pro-



100 YEARS AGO

The very remarkable description of the “Fire Walk” collected by Mr. Andrew Lang and others had aroused a curiosity in me to witness the original ceremony, which I have lately been able to gratify in a visit to Tahiti. I had heard that it was performed in Tahiti in 1897, and several persons there assured me of their having seen it, and one of them of his having walked through the fire himself under the guidance of the priest, Papa-Ita... who had also performed the rite at the island of Hawaii some time in the present year, of which circumstantial newspaper accounts were given... According to these, a pit was dug in which large stones were heated *red hot* by a fire which had been burning many hours. The upper stones were pushed away just before the ceremony, so as to leave the lower stones to tread upon, and over these, “glowing red hot” (according to the newspaper accounts), Papa-Ita had walked with naked feet... I could not doubt that if all these were verified by my own observation, it would mean nothing less to me than a departure from the customary order of Nature, and something very well worth seeing indeed.

From *Nature* 22 August 1901.

50 YEARS AGO

Many of the small plankton animals in the sea which are important as the food of fish such as herring, sprat and mackerel swim upwards towards the surface in the evening and down again to deeper levels after dawn. It is of interest from a purely biological point of view to find out what are the factors which govern these movements, and may also be useful in reaching a better understanding of the shoaling of herring. The plankton consist of small animals of many different kinds, small jellyfish, worms and molluscs, hosts of small crustaceans and many others; they nearly all show this nightly vertical migration upwards. Since it has been developed in so many different groups of animals and must use up so much energy every day — some of them climbing more than a hundred feet — it must clearly be of profound significance in their lives. We do not yet understand its meaning and are still only in the stages of studying the actual movements of the animals in relation to different conditions of light, temperature, pressure, etc.

From *Nature* 25 August 1951.

ject came from the deployment of 55 seismometers, which captured information from natural seismic events occurring around the world from April 1997 to July 1999. These data have been used to produce dramatic three-dimensional images of the present distribution of seismic velocity contrasts (Fig. 1). Such contrasts can be used to obtain a picture of the deep structure in the mantle. James *et al.*⁴ have identified high-velocity mantle structures, with seemingly irregular lower surfaces, beneath the Kaapvaal and Zimbabwe cratons and the Limpopo belt (Fig. 1). These structures extend to depths of 250–300 km, and are the buoyant roots of cratons that tend to stabilize them against tectonic activity and keep them in one piece. By contrast, mantle beneath the younger fold belts to the south of the cratons is of lower velocity, and probably contains more iron. If the rugged lower surface of the high-velocity volumes is real, it may reflect chemical erosion (by upwelling fluids) of the mantle formed in the Archaean.

The results of the seismic tomography are also valuable in cross-checking inferences drawn from other techniques. Previous estimates of the greatest depth to the base of the lithosphere beneath the Kaapvaal craton lay in the range 200–230 km, and were based on analysis of mantle xenoliths^{8,9}. Those estimates are probably within the bounds of error of the seismic and xenolith data: the volcanic action that takes the inclusions to the surface does not necessarily 'sample' the lithospheric column reliably, and the tomography shows large variations in lithosphere thickness over short lateral distances.

The tomography also reveals a region of relatively low seismic velocity across the north of the Kaapvaal craton. This anomaly underlies the Bushveld intrusion, which was formed around 2 billion years ago by the upwelling of more than 455,000 km³ of mantle-derived magma and extends westward towards the related but unexposed Molopo Farm intrusion. The results confirm previous assumptions that intrusion of huge magma volumes into the crust has affected the whole lithosphere. Mantle of lower velocity can indicate material that is at a higher temperature than its surroundings. Here, however, it probably reflects modification of the mantle by fluids, which has resulted in iron enrichment and a consequent decrease in seismic velocity and an increase in density⁹.

Other analyses of seismic data⁵ tell us something about the Mohorovičić discontinuity — the seismic boundary between crust and mantle — beneath southern Africa. Above the high-velocity regions it is sharp and relatively shallow (35–40 km); but beneath the Bushveld intrusion and areas of younger crust it is deeper (45–50 km) with a much less defined boundary. This differ-

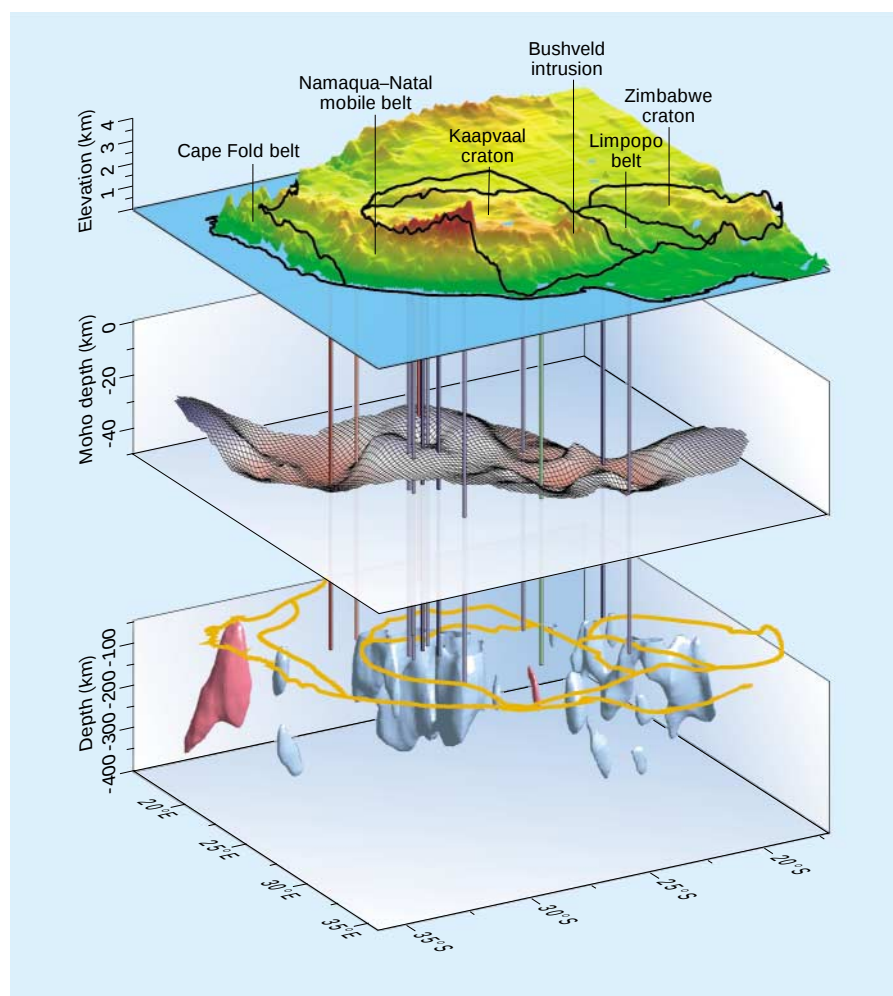


Figure 1 Summary of the results of the Kaapvaal project. **Top**, the topography and principal geological boundaries (black lines) in southern Africa. Vertical lines connect the surface outcrop of some kimberlite magmas to the deep features revealed by interpretation of the seismic data (blue lines, 2.5 Gyr; green, 2–2.5 Gyr; red, < 2 Gyr). **Centre**, the varying depth of the Mohorovičić discontinuity (Moho), which marks the transition between rock types of the crust and those of the mantle. Blue areas indicate the sharp, relatively shallow boundary that exists beneath the cratons; red areas show the deeper and more complex Moho beneath regions where the mantle has been modified by heating or tectonic activity. **Bottom**, cratonic cores (blue) reveal themselves as regions of high seismic velocity lying beneath the Kaapvaal and Zimbabwe cratons and the Limpopo belt. Regions of lower velocity (red) underlie the southern fold belts and the Bushveld complex, which consists of a huge magmatic intrusion. The yellow outline shows the surface geological boundaries. (Image by M. Fouch and D. James; see ref. 12.)

ence could indicate that processes of crustal formation were different in Archaean and post-Archaean times. Or it could be that more complex crust–mantle boundaries are produced by later magmatic or tectonic modification.

Three of the new papers^{3,6,7} describe the integration of different ways of defining the seismic anisotropy of mantle beneath southern Africa. Such studies are significant because they may provide evidence of the deformation that has affected different tectonic regions at different times. The Archaean mantle of cratons shows weaker anisotropy than the adjacent, younger regions. Overall, the authors conclude that the Archaean did not have the large-scale regional fabrics (the 'grain' of deformational features) associated

with the convergent tectonics that characterized later times.

A crucial part of the puzzle in this research is dating the time of lithosphere stabilization — that is, when the crust and the mantle became welded together. The conclusion¹ from whole-rock rhenium–osmium dating of xenoliths is that the south-eastern Kaapvaal craton was stabilized about 2.8 billion years ago. This age is similar to that inferred for xenoliths from the lower crust, and is within the range of previous data on xenoliths from the Kaapvaal mantle¹⁰. However, the *in situ* rhenium–osmium analysis of sulphide minerals in such xenoliths¹¹ has shown that a single xenolith may contain several generations of sulphides, resulting in mixed ages. In the future, this *in situ*

approach may provide more precise ages for lithosphere stabilization, and may distinguish stabilization from later events of crust–mantle disturbance, such as the vertical migration of melts.

Another piece of the puzzle for reconstructing craton formation has come from xenoliths of basalt composition called eclogites, which form a small percentage of the mantle and contain some types of diamonds². These eclogites (and their diamonds) are thought to have formed after craton stabilization, either by intrusion of magmas or by recycling into the mantle of basaltic crust that formed earlier. The stable isotope compositions (including carbon and oxygen) of these eclogite xenoliths support a recycled origin, as the values obtained are similar to those in altered basalts from the sea floor. However, rhenium–osmium ages for the formation of these eclogites are about 500 million years older than the (presumably) surrounding mantle rocks¹. This surprising result may be an artefact of the analytical technique, which may give mixed ages for mantle rocks rather than the age of their original formation, as mentioned above.

There is much speculation about how cratons formed and evolved. A commonly held view is that the processes operating deep in the Earth to form the lithosphere in Archaean times were fundamentally different from those forming the lithosphere today. The Earth's mantle was much hotter in the Archaean, and this caused large-scale melting, with relatively iron-rich magmas adding to the crust and the complementary

magnesium-rich residue forming the buoyant mantle layer of the Archaean lithosphere. There is evidence⁹ from distinctive layered mantle lithosphere in some cratons that large mantle plumes (upwelling blobs of very hot, deep mantle material) provided not only the source of heat for this melting, but also the material that separated into the magmas and residue that formed Archaean cratons.

The Kaapvaal project has provided an unprecedented seismic context for xenolith studies of cratons in southern Africa. Such a context is needed to complement the understanding gleaned from xenoliths in other cratonic regions, such as Siberia, and in the Slave Craton in Canada.

Suzanne Y. O'Reilly is in the GEMOC National Key Centre in the Department of Earth and Planetary Sciences, Macquarie University, Sydney, New South Wales 2109, Australia.

e-mail: soreilly@mq.edu.au

1. Irvine, G. J., Pearson, D. G. & Carlson, R. W. *Geophys. Res. Lett.* **28**, 2505–2508 (2001).
2. Shirey, S. B. *et al. Geophys. Res. Lett.* **28**, 2509–2512 (2001).
3. Ben-Ismail, W., Barruol, G. & Mainprice, D. *Geophys. Res. Lett.* **28**, 2497–2500 (2001).
4. James, D. E. *et al. Geophys. Res. Lett.* **28**, 2485–2488 (2001).
5. Nguuri, T. K. *et al. Geophys. Res. Lett.* **28**, 2501–2504 (2001).
6. Freybourger, M. *et al. Geophys. Res. Lett.* **28**, 2489–2492 (2001).
7. Silver, P. G. *et al. Geophys. Res. Lett.* **28**, 2493–2496 (2001).
8. Boyd, F. R. & Gurney, J. J. *Science* **232**, 472–477 (1986).
9. Griffin, W. L., O'Reilly, S. Y. & Ryan, C. G. in *Mantle Petrology: Field Observations and High-pressure Experimentation: A Tribute to Francis R. (Joe) Boyd* (eds Fei, Y., Bertka, C. M. & Mysen, B. O.) 13–43 (Geochem. Soc., London, 1999).
10. Carlson, R.W. *et al. Proc. 7th Int. Kimberlite Conf. Cape Town*, Vol. 1, 99–108 (Red Roof Design, Cape Town, 1999).
11. Pearson, N. J., Alard, O., Griffin, W. L., Jackson, S. E. & O'Reilly, S. Y. *Geochim. Cosmochim. Acta* (in the press).
12. http://www.agu.org/GRL/images/vol28/g1_28_13L.jpg

Developmental biology

Clocks and Hox

Clifford J. Tabin and Randy L. Johnson

Segmentation is a key feature of many animals. New molecular studies add to our understanding of how vertebrate segments form and how this process is linked to the genes that make each segment unique.

In vertebrates, the spine, ribcage and breastbone are derived from repeated blocks of tissue that begin as identical units in early development and are then modified into unique shapes with different purposes. Some segments, for example, allow the head to move; some are sites of attachment for the muscles involved in breathing; and some protect the organs in the chest. To produce such a body plan, there must be mechanisms both for generating the segments and for giving each its distinct identity. For vertebrates, the task of producing repeated units seems to be controlled partly by a molecular clock in the unsegmented paraxial mesoderm — the tissue from which the units arise. The identity of the units is controlled by the differential expression of genes known

as Hox genes in a nested pattern from the head to the tail¹. Writing in *Cell*, Dubrulle *et al.*² and Zákány *et al.*³ suggest that these processes are causally connected.

Simply put, segmentation in vertebrate embryos occurs as follows. On each side of the neural tube (which forms the spinal cord) is a strip of unsegmented ('presomitic') mesoderm. Cells from this tissue progressively bud off, contributing to somites — the units of cells that will later develop into vertebrae and associated muscles. This differentiation process occurs in a wave that moves gradually from the head to the tail (that is, down the anterior–posterior axis), with presomitic mesoderm in front of the wave and somites in its wake.

The molecular and genetic mechanisms

that control vertebrate segmentation are not yet understood, but emerging evidence supports a long-standing theory known as the 'clock-and-wavefront' model⁴ (Fig. 1). In this model, an autonomous developmental timer (the segmentation clock) interacts with a molecular wavefront of differentiation, which converts information from the clock into spatial information. The model has received support from the discovery of several genes whose expression patterns oscillate in the presomitic mesoderm with the same periodicity as that of somite formation (reviewed in ref. 5).

Dubrulle *et al.*² now provide evidence that the wavefront may correspond to a sharp gradient of fibroblast growth factor-8 (FGF-8) protein within the presomitic mesoderm. Grafting experiments in chick embryos² showed that presumptive somites –I to –V (see Fig. 1) are fixed with respect to their anterior–posterior polarity and boundaries. But the tissue posterior to somite –V is not yet fixed in this way. Dubrulle *et al.* show that this 'undetermined' zone of the presomitic mesoderm corresponds to a posterior domain of high FGF-8 expression.

Dubrulle *et al.* also found that widespread, forced expression of FGF-8 is sufficient to keep the cells of the presomitic mesoderm in an immature, undifferentiated state. Moreover, when FGF-8 protein was applied locally to one part of the presomitic mesoderm, a series of small somites formed. However, the total number of somites remained the same, because a larger-than-normal somite developed posterior to the smaller ones. Conversely, when signalling from FGF-8 was blocked, larger somites formed. How can these results be explained?

The authors' detailed observations² of oscillating gene expression show that forced FGF-8 signalling does not affect the segmentation clock itself. Rather, their data suggest that an interaction between FGF-8 signalling and the clock controls somite size, possibly by specifying the location of boundaries between presumptive somites. So, the clock determines when the boundaries form, and the gradient of FGF-8 determines where. When more FGF-8 is applied to the embryo, the FGF-8 gradient continues further than normal in the anterior direction, and more presomitic mesodermal cells are prevented from contributing to somites. But each somite is programmed to develop at the same time as usual, so each somite ends up containing fewer cells and is therefore smaller. Conversely, blocking FGF signalling shifts the determination front in the posterior direction, and so more cells are allocated to each somite behind the front. This interpretation is consistent with a role for FGF signalling in mediating the wavefront component of the clock-and-wavefront model (Fig. 1).

Once somites have formed, they must

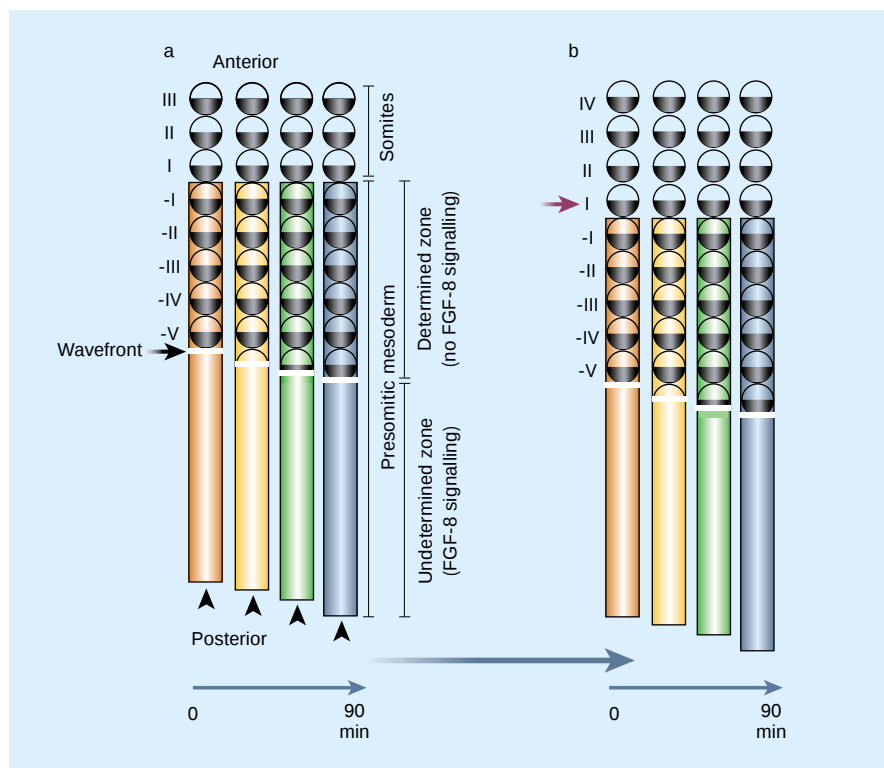


Figure 1 The 'clock-and-wavefront model' of segmentation and patterning in vertebrates. Each vertebra is derived from the fusion of derivatives of an anterior half-somite (transparent semicircles) and of a posterior half-somite (black semicircles). The somites themselves are formed from presomitic mesoderm, in which gene expression oscillates, producing a segmentation 'clock'. The presomitic mesoderm is maintained at a constant length by addition of cells at its posterior end. **a**, The boundary between somites is specified by the interaction of a wavefront of differentiation with the presomitic mesoderm at the start of a new clock cycle (red). This interaction specifies the mesodermal cells as anterior in the first half of the cycle (red and yellow) and posterior in the second half (green and blue). **b**, At the same time as a new cycle begins, a new somite (purple arrowhead) is formed at the anterior end of the presomitic mesoderm. The oscillation of gene expression is regulated, at least in part, by components of the Notch signalling pathway. Signalling from fibroblast growth factor-8 (FGF-8) sets the location of the undetermined zone; the wavefront corresponds to the boundary between cells that express FGF-8 and cells that do not².

express the appropriate complement of Hox genes, which determine the nature of the somites. The ability to manipulate somite size, and hence the number of somites in a given region, allowed Dubrulle *et al.* to investigate whether the boundaries of Hox-gene expression follow somite number or absolute anterior-posterior position. The answer is that the embryo seems to count somites to set the boundaries of Hox expression.

Zákány *et al.*³ suggest that the coordination of Hox-gene expression with segmentation requires the molecules of the segmentation clock. They found that several Hox genes are activated in the presomitic mesoderm in mice at the location where the next somite will form. This region coincides both spatially and temporally with a wave of expression of *lunatic-fringe*, a gene that is cyclically expressed in presomitic mesoderm and is controlled by signalling from Notch—a component of the segmentation clock. So it seemed that Hox expression might likewise be regulated by the clock. The authors

confirmed this by showing that mice with a mutation in a protein downstream of Notch fail to activate significant levels of Hox expression in the presomitic mesoderm. At minimum, this provides a mechanism for aligning the levels of Hox expression with the

segments, as Hox-gene expression is triggered up to the sharp boundary formed by the action of the clock and wavefront. Once a particular Hox gene is activated by the clock, its anterior boundary of expression is maintained at the level of the somite that was formed at the time the gene was switched on.

These studies^{2,3} provide convincing evidence of a link between activation of Hox genes and segmentation. It seems that once Hox genes are activated by the clock in the presomitic mesoderm, they maintain their expression boundaries in somites and their derivatives. So the time at which a Hox gene is activated establishes its boundary of expression in the embryo. But it is not clear why one Hox gene is activated rather than another in a given clock cycle. As there are more somites than Hox genes, the mechanism must be more complicated than 'one somite, one Hox gene'. The activation of Hox genes follows the physical order of the genes on the relevant chromosome, and this 'temporal co-linearity' depends on the release of the DNA containing the Hox genes from a repressive configuration⁶. The data discussed here suggest that Notch signalling is involved in periodic activation of the expression of unrepressed Hox genes. The details of how the release from repression is coordinated with the segmentation clock will be the next piece of the puzzle. ■

Clifford J. Tabin is in the Department of Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA.

e-mail: tabin@rascal.med.harvard.edu

Randy L. Johnson is in the Department of Biochemistry and Molecular Biology, University of Texas, MD Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, Texas 77030, USA. e-mail: rjohnson@odin.mdacc.tmc.edu

1. Krumlauf, R. *Cell* **78**, 191–201 (2001).
2. Dubrulle, J., McGrew, M. J. & Pourquié, O. *Cell* **106**, 219–232 (2001).
3. Zákány, J., Kmita, M., Alarcon, P., de la Pompa, J.-L. & Duboule, D. *Cell* **106**, 207–217 (2001).
4. Cooke, J. & Zeeman, E. C. *J. Theor. Biol.* **58**, 455–476 (1976).
5. Pourquié, O. *Curr. Top. Dev. Biol.* **47**, 81–105 (2000).
6. Kondo, T. & Duboulet, D. *Cell* **97**, 407–417 (1999).

Planetary science

Gases make a rare appearance

Typhoon Lee

The unexpected discovery of noble gases within silicate pockets in a primitive meteorite is hard to explain. Could they have been captured from the solar wind streaming from the young Sun?

The five noble gases — helium, neon, krypton, argon and xenon — are rather aloof. They neither react with other elements to form solids nor condense easily. Although they are quite abundant in the Universe, their strong preference to remain gaseous means that their concentrations in

rocks are very low, so they are also known as rare gases. But on page 795 of this issue, Okazaki and co-workers¹ report surprisingly high concentrations of noble gases in the silicate beads found in a primitive meteorite. The most primitive, but also the most common, meteorites are chondrites, so called

because they contain chondrules — silicate spherules solidified from melt droplets formed by a mysterious heating event early in the history of the Solar System. The discovery of large amounts of volatile noble gases in these chondrules is unexpected because it is widely believed either that the gases never existed in the material that formed the chondrules or, if they were present, that they were driven out by the heating event. Although noble gases have not yet been found in chondrules in any other meteorites, this discovery may lead to new insight into the processes that formed our Solar System.

This is not the first time that rare gases have been used to probe the history of the Solar System. The addition of even a tiny amount of a noble gas to a solid causes a large fractional increase, making it easier to detect. So the noble gases provide an extremely sensitive record of processes that may be too subtle to detect by other means. For example, the extinct radioactive nuclide I-129, which existed for only a short time in the early Solar System, was found through the detection of its daughter product, Xe-129. Another example is that, when a rock is irradiated by cosmic rays, tiny amounts of magnesium and silicon are converted into Ne-21. So the accumulation of Ne-21 in a meteorite is a reliable measure of the time it took to travel from its parent asteroid to the impact with Earth. The detection of rare gases in chondrules is perhaps the most unexpected of these discoveries, and may eventually lead to a better understanding of the formation and evolution of chondrites.

The chondrules studied by Okazaki and co-workers¹ came from meteorite Yamato-791790 (hereafter referred to as Y-79), an enstatite chondrite found in Antarctica. The authors used pulses from a fine laser beam to heat small amounts of material — typically 0.7 micrograms — from different areas of Y-79. Rare gases released from each area were analysed in a mass spectrometer. Test areas inside the chondrules released about 80, 0.4 and 0.08 million atoms of Ar-36, Kr-84 and Xe-132, respectively. In contrast, the amount of rare gases released from the matrix between the chondrules was ten times lower and approaches the instrumental background levels. Together these data indicate that rare gases in Y-79 reside primarily in chondrules.

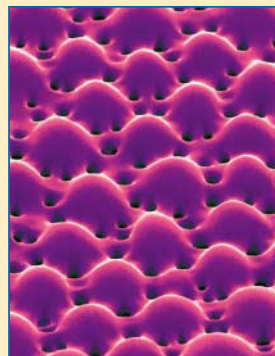
How do noble gases get into meteorites? One possible explanation comes from studies of samples from the Moon. Without the protection of either a magnetic field or an atmosphere, the surface of the Moon is constantly being bombarded by the solar wind. Travelling at a speed of around 500 km s⁻¹, this outflow of ionized gas from the Sun directly impinges upon the lunar surface, and the fast-moving particles penetrate into the solid before stopping within the first micrometre or so. Some of these ionized

Optics

Armed for light sensing

Light is a primary driving force in nature, and most organisms have some type of light-detection system, often one that uses lenses. But it comes as a surprise to learn from Joanna Aizenberg and colleagues, writing elsewhere in this issue (*Nature* **412**, 819–822; 2001), that a species of brittlestar, *Ophiocoma wendtii*, possesses a remarkable microlens array. Brittlestars belong to the group of marine invertebrates known as echinoderms, and Aizenberg *et al.* find that the calcium carbonate (calcite) that makes up the external skeleton of *O. wendtii* also forms light-sensing arrays.

The arrays are found in the skeletal plates that protect the upper-arm joints on each of the brittlestar's five arms. In some light-insensitive species, these plates constitute a relatively open, three-dimensional mesh of single-crystal calcite, with mesh pore sizes of around one-hundredth of a millimetre. But in *O. wendtii* the outer surface of this mesh has a characteristic array of larger, spherical protuberances, each about one-twentieth of a millimetre in diameter and linked to six neighbours (see the figure). When seen in cross-section, each protuberance has the



appearance of a double lens: the radius of curvature of the upper face is about 20 to 30 μm , that of the lower face rather less. This combination gives a focal length of about 10 μm , with a focal-spot size of less than 3 μm . There are bundles of nerve fibres of about that size at each focal spot, and the authors suggest that these bundles are responsible for the documented sensitivity of *O. wendtii* to light stimuli.

The construction and operation of microlenses have strict requirements. First, there has to be exquisite control of calcite growth to form the lens structures. Second, calcite is optically anisotropic, with different refractive indices for light polarized in different directions. So, to avoid birefringence effects, it has to grow as single crystals

with the optical axis parallel to the axis of the double lens. Third, each microlens should ideally have minimal optical aberration, and that seems to be the case. The authors have checked this last point both in experiments using an extracted array of lenses as the focusing elements and by modelling the optical response of such structures.

Human ingenuity came up with microlens arrays only a few years ago, and they are used in directional displays and in micro-optics, for example as signal-routing connectors for signal processing. Once again we find that nature foreshadowed our technical developments. The same applies to photonic solids, structures that can selectively reflect light in all directions. Photonic materials have stimulated much research over the past ten years because of their potential in light manipulation, yet they are to be found in opals and in the wings of butterflies. But then, nature has been in the business of developing functioning optical structures for a very long time.

Roy Sambles

Roy Sambles is in the Department of Physics, University of Exeter, Stocker Road, Exeter EX4 4QL, UK. e-mail: j.r.sambles@exeter.ac.uk

particles are noble gases. When lunar soil samples are heated in a laboratory vacuum, the implanted solar wind particles can diffuse back out and be measured. This is how the 'solar' pattern of rare gases was determined² (Fig. 1, overleaf). This pattern favours the lighter rare gases, so the abundances drop exponentially from helium to xenon.

Rare gases found in meteorites follow one of two patterns. One of them is reminiscent of that found in the Earth's atmosphere (Fig. 1), so it is termed 'planetary', even though its origin is still unknown. The other type closely follows the solar pattern determined from lunar soil samples, and so presumably came from a similar process of solar wind implantation, except that the exposure took place on the surface of asteroids instead of the Moon. Until now, most

of the rare gas in meteorites was found in fine-grained fragments attributed to the ancient surface layer of asteroids.

In the case of Y-79, the proportions of rare gases released from the chondrules were roughly 2,000:5:1 for Ar-36, Kr-84 and Xe-132. The pattern of high Ar:Xe but moderate Kr:Xe seems to indicate that the rare gases in Y-79 follow the solar pattern. This makes solar-wind implantation an attractive hypothesis. But no trace of helium or neon was detected in Y-79 and the argon seems to have been depleted relative to what might be expected for the solar pattern based on krypton and xenon alone (Fig. 1). There is, however, an apparent correlation between the missing gases and their masses. The lightest gases, helium and neon, are completely depleted, whereas argon is only partially

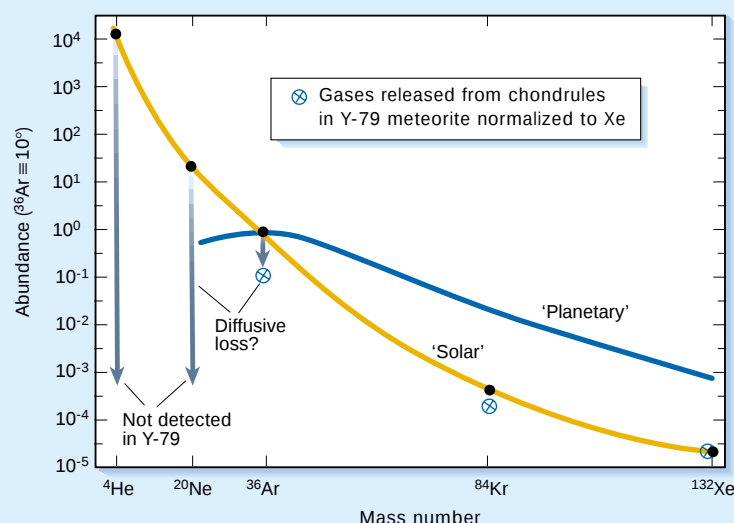


Figure 1 Solar and planetary patterns of rare-gas abundances (modified from Wasson⁴). The 'solar' pattern arises when solar-wind particles are captured in the surface layer of lunar grains. The 'planetary' pattern resembles that of the terrestrial atmosphere. The rare gases found by Okazaki *et al.*¹ in chondrules of the Yamato-791790 enstatite meteorite are close to the solar pattern for argon, krypton and xenon. The lack of neon and helium and the slight depletion in argon (relative to krypton and xenon) are likely to be due to diffusion processes at high temperatures that preferentially remove light rare gases. The unexpected discovery of noble gases in these chondrules has implications for models of chondrule formation.

depleted, and the heaviest gases are intact. The most common process to produce a mass-dependent loss of gases is thermal diffusion, which preferentially allows the light gases to escape. So the most plausible interpretation of the pattern of noble gases found in Y-79, as Okazaki *et al.* suggest, is that rare gases were implanted when the chondrules were exposed to the solar wind, but the lighter elements were subsequently lost through diffusion. For this to occur, the chondrules must have been heated to high temperatures.

If this explanation is correct then it offers interesting clues to the nature of the flash heating event that produced the chondrules in the first place. The origin of this heating is a century-old mystery. During their early formative years, stars like the Sun are surrounded by an accretion disk of gas and dust (the disk for our own planetary system is called the solar nebula) that adds directly to the growth of the young star. This period of growth is characterized by a very active Sun and hence a strong solar wind. But solar-wind implantation of rare gases on the surfaces of asteroids could not have taken place until the solar nebula had been dissipated (either blown away or condensed into planets). Otherwise, the asteroid belt would have been shielded from the solar wind because energetic particles cannot penetrate very far through the solar nebula.

The lack of rare gases in the matrix between the chondrules in Y-79 suggests that exposure of the chondrules to the solar wind took place before they coalesced to form larger aggregates and eventually asteroids.

So these chondrules captured the solar wind at an earlier time than is possible for the fine-grained material that was exposed after the asteroids were assembled. Moreover, to be exposed to the solar wind at this earlier epoch, the chondrules in Y-79 had to be located near or outside the surface of the solar nebula. So this new result does not support models of chondrule formation that suggest they formed in the highly shielded mid-plane of the accretion disk, at distances corresponding to the current position of the asteroid belt.

In an alternative theory, known as the X-wind model³, the interaction of the accretion disk with the young Sun's magnetic field produced flares of highly energetic particles, which may have been responsible for the initial flash heating of the chondrules. In this model, chondrules were generated near the inner edge of the accretion disk where the temperatures were high and the young solar wind was strong. The solar type of rare gases found in the chondrules of Y-79 are so readily explained by the X-wind model that it is a puzzle why so few of these gas-rich chondrules have been found. No doubt some meteoritists will now be looking more closely. ■

Typhoon Lee is at the Institute of Earth Sciences, Academia Sinica, Box 1-55, Nankang, Taipei, Taiwan.

e-mail: typhoon@gate.sinica.edu.tw

1. Okazaki, R., Takaoka, N., Nagao, K., Sekiya, M. & Nakamura, T. *Nature* **412**, 795–798 (2001).
2. Ozima, M. & Podosek, F. A. *Noble Gas Geochemistry* (Cambridge Univ. Press, 1983).
3. Shu, F. H., Shang, H., Gounelle, M., Glassgold, A. E. & Lee, T. *Astrophys. J.* **548**, 1029–1050 (2001).
4. Wasson, J. T. *Meteorites* (Springer, Berlin, 1974).

Daedalus

Neutrinos in orbit

The neutrino telescope, essentially 100,000 gallons of cleaning fluid at the bottom of a gold mine, arouses Daedalus's strong admiration: he would so like to have invented something like that himself. He now wants to make it directional, like a normal telescope. He plans a tube many metres long, underground as before, and full of dry-cleaning fluid, or another particle-detecting liquid. Only in the direction of the steerable tube, when neutrinos traverse it from end to end, is it at all sensitive. Several parallel tubes even permit energy-dispersion studies.

Solar neutrinos do not quite ignore the Sun and planets. Unlike photons, they are now known to have mass. They must be slowed by solid objects, perhaps even down to Solar or planetary orbital speeds. They could even be deviated by asymmetric encounters with matter, and so enter orbit. There could be quite a lot of neutrinos in nearby space, slowed by deep encounters with Mercury, Venus, the Sun, Earth and the Moon. They will be orbiting these solid bodies, entering them with little loss per orbit, and be topped up by new arrivals. Their equilibrium concentration will reflect the deep neutrino absorption of these objects.

Neutrinos slowed and deviated by the Moon, for example, should enter an Earth orbit from which a few of them would penetrate below the surface once per orbit, and traverse the telescope. Neutrinos slowed by the nearer planets but little deviated by them should enter a narrow elliptical Solar orbit. Some of them should traverse the telescope twice per orbit. At the other extreme, they would swing round the outer layers of the Sun, sampling its own neutrino-absorbing and -deviating properties. Daedalus's new telescope will look for these wayward particles. He hopes to calculate those orbits best configured to encounter new orbital neutrinos.

Of course, the neutrino flux of any orbit must carry the 'signature' of the whole of that orbit. Neutrinos deviated by the Moon, and passing through it once per orbit, will have an orbital equilibrium concentration reflecting its deep neutrino-absorbing and -deviating properties. These could be calculated, and would slowly give reliable information about the deep geology of the Moon. Similarly, the deep geology of the planets, and even of the Sun, would gradually become clear. It should reveal how opaque, or how deviating, these bodies are to entering neutrinos.

David Jones

Export of organic carbon from peat soils

Warmer conditions may be to blame for the exodus of peatland carbon to the oceans.

We have observed a 65% increase in the dissolved organic carbon (DOC) concentration in freshwater draining from upland catchments in the United Kingdom over the past 12 years. Here we show that rising temperatures may drive this process by stimulating the export of DOC from peatlands. Our results indicate that the flux of aged, riverine DOC of terrestrial origin, now recognized as a significant supplier of DOC to oceans¹, may increase substantially as a result of global warming.

DOC concentrations have increased significantly ($P < 0.05$) at 20 of 22 sites in the UK Acid Waters Monitoring Network², according to the Seasonal Kendall trend analysis (Fig. 1a). These sites span a wide range of acid-deposition levels, soils, topographies, land uses and geographical locations. Annual increases, averaging 5.4%, are proportional to mean DOC concentration ($R^2 = 0.81$, $P < 0.001$). As freshwater DOC concentrations are linked to storage of carbon in catchment soil³, this indicates that increases are driven by regionally consistent processes within this carbon store, and that they are greatest at sites with large stores of soil carbon, such as peatlands (Fig. 2).

Although an inverse relationship has been proposed between mineral acidity and the generation of DOC⁴, we observed similar proportional increases in DOC at remote, unacidified sites, as well as at those

recovering from anthropogenic acidification. Changes in land use or river discharge do not account for the observed increases. However, the Central England Temperature Record⁵ shows that mean temperatures were 0.66 °C higher in the 1990s than in the three preceding decades, and this factor could have influenced all sites.

The enzyme phenol oxidase has been proposed to regulate carbon storage in peatlands⁶. We therefore studied the thermal responses of peatland phenol oxidase in relation to the export of DOC. We subjected peat soil to a thermal gradient of 2–20 °C. Phenol oxidase activity was greater at higher temperatures, although this enzyme is known to be highly constrained in these waterlogged soils. An increase of 10 °C led to a 36% increase in activity ($Q_{10} = 1.36$). This was accompanied by an equivalent increase in DOC release ($Q_{10} = 1.33$) and an even greater increase in release of phenolic compounds ($Q_{10} = 1.72$) from the soil matrix. This selective enrichment with phenolic compounds (Fig. 1b) is noteworthy because of the inhibitory character of these compounds⁷. Under warmer conditions, selective enrichment should impair the metabolism of the remaining DOC⁸, allowing even more DOC to reach the oceans.

Pre-aged terrestrial sources of carbon are important contributors to the oceanic carbon budget¹. Peat-accumulating wetlands have created a significant terrestrial store⁹ of such highly aged organic matter, despite exporting more organic carbon per unit area than any other significant biogeographical land type in the world¹⁰. We have shown that a key terrestrial carbon store could currently be being relocated to the oceans. The rate of movement is likely to increase further if global temperatures increase. Investigation of the fate of that material in the recipient ecosystem will be important.

C. Freeman*, C. D. Evans†, D. T. Monteith‡, B. Reynolds§, N. Fenner*

Resource availability

Ancient homes for hard-up hermit crabs

Mollusc shells are a vital but sometimes scarce resource for hermit crabs, protecting them from mechanical damage and desiccation, but they require continual replacement as the crab grows. I have discovered that *Coenobita rugosus*, a large, tropical, semi-terrestrial hermit crab, will resort to using fossil shells when no other suitable casing is available.



Figure 2 Peatland bog: northern peatlands remove carbon dioxide from the atmosphere faster than it is released, so they now contain 20–30% of the world's soil carbon stock. But this may be changing in response to warmer conditions.

B. Reynolds§, N. Fenner*

*School of Biological Sciences, University of Wales, Bangor LL57 2UW, UK

e-mail: c.freeman@bangor.ac.uk

†Centre for Ecology and Hydrology, Wallingford OX10 3BB, UK

‡Environmental Change Research Centre, University College London, London WC1H 0AP, UK

§Centre for Ecology and Hydrology, Bangor LL57 2UP, UK

1. Raymond, P. A. & Bauer, J. E. *Nature* **409**, 497–500 (2001).
2. Monteith, D. T. & Evans, C. D. *The United Kingdom Acid Waters Monitoring Network, 10 Year Report* (ENSIS, London, 2000).
3. Hope, D., Billett, M. F., Milne, R. & Brown, T. W. *Hydrol. Process.* **11**, 325–344 (1997).
4. Krug, E. C. & Frink, C. R. *Science* **221**, 520–525 (1983).
5. Parker, D. E., Legg, T. P. & Folland, C. K. *Int. J. Climatol.* **12**, 317–342 (1992).
6. Freeman, C., Ostle, N. & Kang, H. *Nature* **409**, 149 (2001).
7. Wetzel, R. G. *Hydrobiologia* **229**, 181–198 (1992).
8. Freeman, C., Lock, M. A., Marxsen, J. & Jones, S. E. *Freshwat. Biol.* **24**, 159–166 (1990).
9. Gorham, E. *Ecol. Appl.* **1**, 182–195 (1991).
10. Lugo, A. E., Brown, S. & Brinson, M. M. in *Ecosystems of the World* vol. 15 (eds Lugo, A. E., Brown, S. & Brinson, M. M.) 53–85 (Elsevier, Amsterdam, 1989).

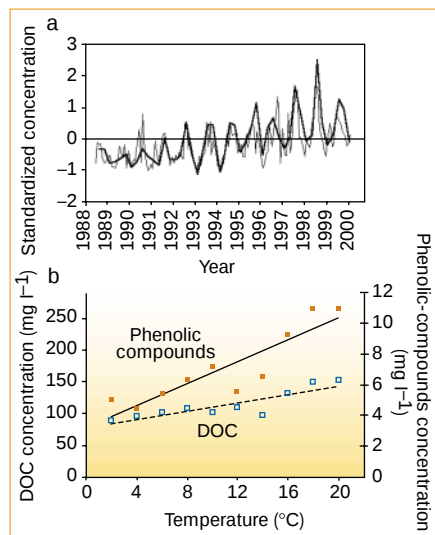


Figure 1 Changing concentrations of dissolved organic carbon (DOC). **a**, Time series of median standardized DOC concentrations determined from quarterly data for 11 lakes (thick line) and monthly data for 11 streams (thin line) in the UK Acid Waters Monitoring Network (standardized concentrations for each site have a mean of zero and a standard deviation of one). **b**, Laboratory observations of increased concentrations of DOC and phenolic compounds in peat soil in response to rising temperature.

These unlikely mobile homes fall out of coastal limestone as it is eroded by the sea in southwestern Madagascar, placing the occupants alongside *Homo sapiens* as resourceful exploiters of prehistoric animal remains.

Hermit crabs procure their protective casing, typically a gastropod mollusc shell, from other hermit crabs^{1,2}, rather than directly from living molluscs³. Reducing predation is thought to be the most important factor in determining shell selection by tropical hermit crabs⁴, but adults of intertidal species living above the high-water

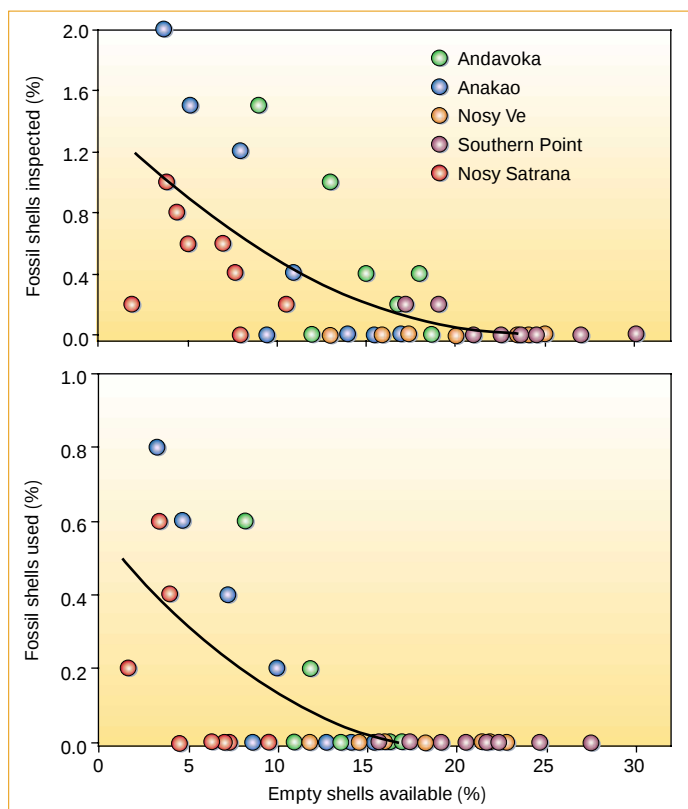
mark also pick them for their anti-desiccation properties⁵. Hermit crabs often suffer from a shortage of shells, however, and so are forced to make do with less suitable shells or even with artificial materials⁶. *Coenobita rugosus* is one species that uses fossil shells as a substitute (Fig. 1) — a much scarcer resource that drops out of limestone rocks as they are eroded.

I found that the number of fossil shells inspected and used by these crabs is proportional to the lack of availability of local shells (referred to as the degree of resource restriction and measured as the proportion of suitable empty shells at three of five sites; Fig. 2). The fossil shells used were from the short-spined, shore-dwelling marine snails *Nassarius*, *Nerita* and *Turbo*. Fossil shells were not used at any locality where suitable empty shells comprised over 13% of all shells; however, when the proportion of these fell below 9%, one or more of 500 crabs examined were found to use fossil shells. One other local hermit-crab species (of seven species investigated), *Calcinus latens*, was seen to use fossil shells (*Nerita*), but only on one occasion. The relationship between the percentage of fossil shells inspected and the restricted availability of non-fossil shells (Fig. 2, top) was similar to that for actual usage of fossil shells and restricted non-fossil-shell availability (Fig. 2, bottom).

Explanations for this behaviour are complex, as the crab is unlikely to appreciate the nature of a shell until it has inspected it. Fossil shells are generally found in close association with the limestone rock of their origin, rather than in locations at which hermit crabs might find many empty shells, such as at the strand line. The relationship between fossil-shell inspection and shell restriction suggests that crabs in shell-restricted areas search for local shells first.

Shelter for hermit crabs could potentially be provided by discarded shells, shells

Figure 2 Percentage of fossil shells inspected and used by the hermit crab *Coenobita rugosus* as shelter in relation to the availability of empty shells at five different sites in southwestern Madagascar. Each point corresponds to 500 inspections by hermit crabs (top) or to usage by 500 crabs (bottom). Fitted curves are second-order regressions with r^2 values of 0.47 and 0.46, respectively; $P < 0.01$ (ANOVA) in both cases. Eight counts of 500 gastropod shells were made in each of 5 supralittoral study sites (listed in upper right). Crabs inhabiting fossil shells were identified (as species), as was the fossil (as genus). Percentages of empty mollusc shells, fossil shells and used shells that were mechanically damaged or blocked were all recorded and are not represented in the graphs. Further details are available from the author.



already occupied by another hermit crab or by a living mollusc, and by buried or fossil shells. Of the fossil shells I examined, 93–100% had severe structural hindrances that would block any potential occupant. But the rarity of exploitation by hermit crabs of fossil shells cannot simply be due to the scarcity of usable fossil shells, because otherwise the extent of usage would be three times higher (average frequency of non-fossil shells is 99% of usable shells and average usage is 99.7%; average frequency of fossil shells is 1% of usable shells and average usage is 0.3%). Despite this, fossil shells are probably the third-largest shell supplier of all the potential shell sources as, in common with observations made elsewhere, no shell excavation^{6,7} or requisition from living molluscs³ was seen.

In southwestern Madagascar, as in many parts of the world, particularly East Africa⁸, collecting of shells on the shore is important in artisan fisheries⁹. Living molluscs are generally removed from their shells at the high-tide mark, where piles of shells, which are mostly too large to be of use to hermit crabs, are deposited. However, as shorelines become more intensively exploited, the mean shell size of molluscs collected in this way is getting smaller^{8,9}. These piles of discarded shells should therefore eventually help to release hermit crabs from the prevailing shell restriction, and render the use of fossil shells unnecessary.

David K. A. Barnes*†‡

*Department of Zoology and Animal Ecology, University College Cork, Lee Maltings, Cork, Ireland

†Frontier, 77 Leonard Street, London EC2A 4QS, UK

‡Present address: British Antarctic Survey, High Cross, Maddingley Road, Cambridge CB3 0ET, UK e-mail: dkab@ucc.ie

1. Reese, E. S. *Anim. Behav.* **10**, 347–360 (1962).
2. Hazlett, B. A. *Behav. Ecol. Sociobiol.* **6**, 177–184 (1980).
3. Rutherford, J. D. *Veliger* **19**, 438–439 (1977).
4. Bertness, M. D. *J. Exp. Mar. Biol. Ecol.* **64**, 159–187 (1982).
5. Burggren, W. W. & McMahon, B. R. *Biology of the Land Crabs* (Cambridge Univ. Press, Cambridge, 1988).
6. Fotheringham, N. *Ecology* **57**, 570–578 (1976).
7. Kellogg, C. W. *J. Exp. Mar. Biol. Ecol.* **22**, 101–111 (1976).
8. Barnes, D. K. A., Corrie, A., Whittington, M., Carvelho, M. A. & Gell, F. J. *Shellfish Res.* **17**, 51–58 (1998).
9. Barnes, D. K. A., Rawlinson, K., Zucco, C., Lawford, L. & Purnell, L. in *Ecology and Biodiversity of Madagascar, a Contribution from Ireland* (eds Barnes, D. K. A. & Sleeman, D. P.) (Irish Biogeogr. Soc., in the press).

correction

Early visual experience and face processing

R. Le Grand, C. J. Mondloch, D. Maurer, H. P. Brent *Nature* **410**, 890 (2001)

There is a mistake in Table 1 and its legend, as the result of a coding error. The final two sentences of the legend should read: “For adults, inversion reduced the accuracy of the configural set significantly more than that of the featural set (significant interaction between orientation and stimulus set ($F_{1,25} = 13.13$, $P < 0.01$), followed by analyses of simple effects ($F_{1,27} = 62.57$, $P < 0.01$)). Patients were less accurate than controls (matched for age, gender, race and handedness) only for the upright configural set (significant interaction between group, orientation and stimulus set ($F_{1,26} = 21.92$, $P < 0.01$)).” Values in the ‘featural, upright’ column should be 80, 89 and 85 for adults, controls and patients, respectively. This error does not affect our conclusions.



Figure 1 Houses for hermit crabs: sometimes very old homes are better than newer ones, particularly in situations where non-fossil shells are in short supply.

Efficiency and ambiguity in an adaptive neural code

Adrienne L. Fairhall, Geoffrey D. Lewen, William Bialek & Robert R. de Ruyter van Steveninck

NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540, USA

We examine the dynamics of a neural code in the context of stimuli whose statistical properties are themselves evolving dynamically. Adaptation to these statistics occurs over a wide range of timescales—from tens of milliseconds to minutes. Rapid components of adaptation serve to optimize the information that action potentials carry about rapid stimulus variations within the local statistical ensemble, while changes in the rate and statistics of action-potential firing encode information about the ensemble itself, thus resolving potential ambiguities. The speed with which information is optimized and ambiguities are resolved approaches the physical limit imposed by statistical sampling and noise.

More than forty years ago it was suggested that neural codes may constitute efficient representations of the sensory world^{1,2}. Within the framework of information theory, efficient coding requires a matching of the coding strategy to the statistics of the input signals³. Recent work shows that the sequences of action potentials from single neurons provide an efficient representation of complex dynamic inputs^{4–7}, that adaptation to the distribution of inputs can occur in real time^{8–10} and that the form of the adaptation can serve to maximize information transmission^{10,11}. However, adaptation involves compromises. An adaptive code is inherently ambiguous: the meaning of a spike or a pattern of spikes depends on context, and resolution of this ambiguity requires that the system additionally encode information about the context itself. In a dynamic environment the context changes in time and there is a tradeoff between tracking rapid changes and optimizing the code for the current context. Here we examine the dynamics of adaptation to statistics in motion-sensitive cells in the visual system of a fly, *Calliphora vicina*, and find that different aspects of adaptation occur on timescales that range from tens of milliseconds to several minutes. The speed of adaptation to a new input distribution must be limited by the need to collect statistics. Adaptation of the neural input/output relation to optimize information transmission approaches this theoretical maximum speed. This rapid adaptation of the input/output relation leaves the longer timescales in the response dynamics as a nearly independent channel for information, resolving potential ambiguities of an adaptive code.

Multiple timescales in statistical adaptation

Codes are defined by the response not just to a particular stimulus, but to a statistical distribution of inputs, which provides the context. Statistical adaptation involves matching the neural code to this input distribution. If the input signals are gaussian the only quantities that can govern statistical adaptation are the mean (as in light–dark adaptation in the retina) and the correlation function. Here we consider signals with a fixed, short correlation time and zero mean, so that the only free statistical parameter is the variance. We record action potentials from an identified motion-sensitive neuron in the fly visual system (H1 (refs 12, 13)). The input signal is the angular velocity of motion across the visual field; a sudden change in variance might mimic what happens as a fly makes a transition from cruising flight¹⁴ ($S_{r.m.s.} \sim 230$ degrees s^{-1}) to acrobatic or chasing behaviour¹⁵ ($S_{r.m.s.} \sim 1,200$ degrees s^{-1}). To track the cell's response to changes in the local stimulus statistics, we modulate a (nearly white) zero-mean signal $z(t)$ in time with a slowly varying amplitude $\sigma(t)$, such that the angular velocity stimulus is given by $S(t) = z(t)\sigma(t)$ (see Methods). Thus the

stimulus has a time-varying distribution that is locally characterized solely by its (varying) standard deviation.

A simple probe of adaptation to statistics is a switching experiment⁸, shown in Fig. 1a, where the standard deviation switches periodically between σ_1 and σ_2 , with a cycle time of T ; we choose a new random sequence $z(t)$ in each cycle. Averaging over many cycles we obtain the spike rate as a function of time, which reflects the dynamics of the response to $\sigma(t)$. As in related experiments^{8,16}, the rate after an abrupt increase in σ first jumps to a high value, then gradually decreases almost to a steady state over the period of the switch (Fig. 1c). It is tempting to identify this transient as the (unique) timescale of adaptation, but as the experiment is repeated with increasing periods T , from 4 to 40 s, the apparent decay time also increases (Fig. 1b). In fact, the decay time scales linearly with the switching period T (Fig. 1c). This scaling holds for the transient after increases but not decreases in variance, suggesting that different mechanisms may be at work in the different directions of adaptation. Similarly, in response to sinusoidal modulations in $\sigma(t)$, the rate is shifted in time with respect to $\sigma(t)$ by a fixed fraction of the modulation period T , for T from 10 to 240 s (data not shown; see ref. 17). The adaptation of the firing rate clearly has access to many timescales¹⁸, and the observed timescale depends on the design of the experiment.

Input/output relations and information transmission

If the spike rate changes when stimuli remain statistically the same—as for adaptation after a switch—the system's rule for generating a spike must be changing; we will quantify this by constructing an input/output relation. Under stationary conditions, the input/output relation of H1 is adapted to the stimulus ensemble in such a way that the stimulus input is rescaled by its standard deviation¹⁰. It is natural to assume that slow transients in the rate dynamics are indicative of or caused by the progress of this adaptation, so we follow the evolution of the input/output relation after a sudden change in statistics, as in the switching experiment.

Even for statistically stationary signals, the spiking probability can be affected by the history of the motion stimulus in a 400-ms window preceding the spike. Thus 'the stimulus' at time t really is a high-dimensional object describing the whole history $S(t)$ for $t_{\text{spike}} - 400 \text{ ms} < t < t_{\text{spike}}$. To study the input/output relation we require a lower-dimensional description of the stimulus, and, as in ref. 10, we do this by linearly projecting or filtering the signal $S(t)$ to obtain a single number s for each spike time (see Methods). The most relevant projection is a smoothed version of the velocity. Although there are interesting adaptations in these linear filters that are analogous to those previously observed^{9,19,20}—possibly also

implicated in information maximization, as has been suggested in other cases^{9,21}—our focus here is on the nonlinear relation between the spike-firing probability and the smoothed velocity signal.

The measured input/output relations are shown in Fig. 2 for selected points in the cycle. When measured in physical units, the input/output relations in the two halves of the cycle differ strongly: when the dynamic range of the signal is large, significant modulations in firing probability require proportionately large variations in the signal; when the dynamic range is small, a much smaller change in input evokes the same modulation in output. The effect is large: under different adaptation conditions the firing probabilities in

response to the same input can differ by orders of magnitude. Thus, the input/output relation is not a static property of the system, but adapts, dynamically, so that the input seems to be measured in units proportional to the local standard deviation.

Strikingly, the differences in the input/output relation are established within the first second after the switch, by which point the input/output relation (and the locally computed linear filter; data not shown) is already very similar to the steady-state form (Fig. 2). More precisely, the transient and steady-state input/output relations differ simply by a scale factor in the spike rate, as seen by comparison with the curves shown in normalized units, where the transient and steady-state curves collapse. This scale factor corresponds to an overall modulation by the time-varying spike rate—the scaling with respect to the input remains the same. Thus the

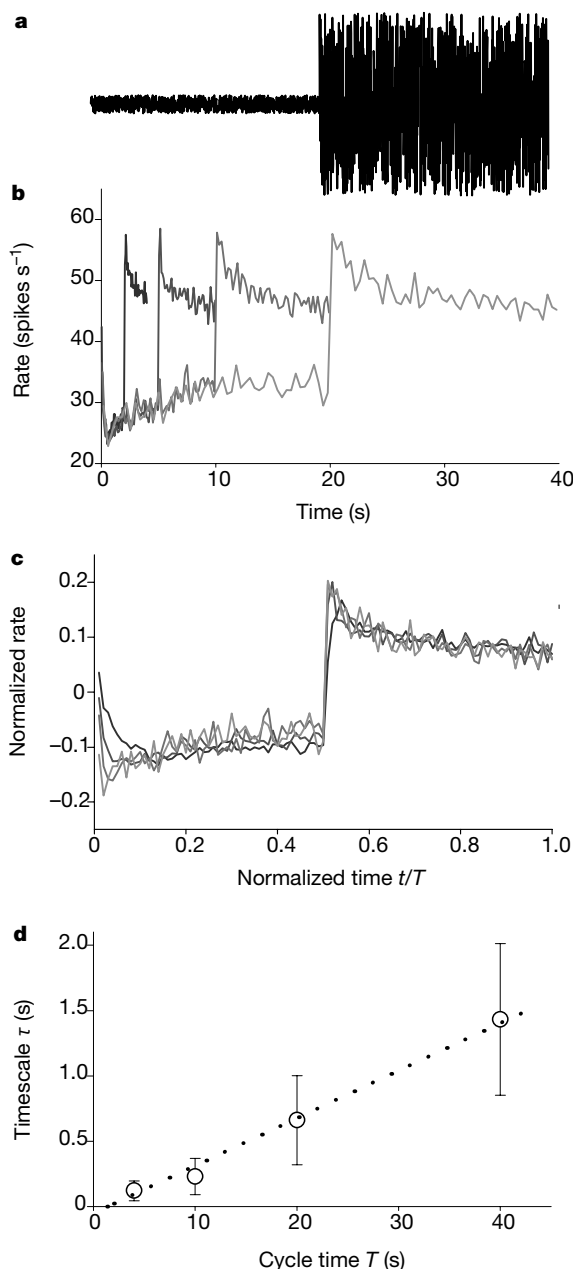


Figure 1 Variance-switching experiment. **a**, The stimulus is a white-noise velocity signal modulated by a square wave envelope that switches between two values, σ_1 and σ_2 , with period T . **b**, Averaged rate as a function of time relative to the cycle, for $T = (4, 10, 20, 40)$ s, represented by a progressively lighter grey scale. **c**, The same curves as in **b** but with the time axis normalized by the period T and with the rate normalized to a mean of zero and unit standard deviation. **d**, The decay timescale τ , extracted from the curves of **a** by fitting an exponential to the decay of spike rate after an upward switch in variance, plotted as a function of T .

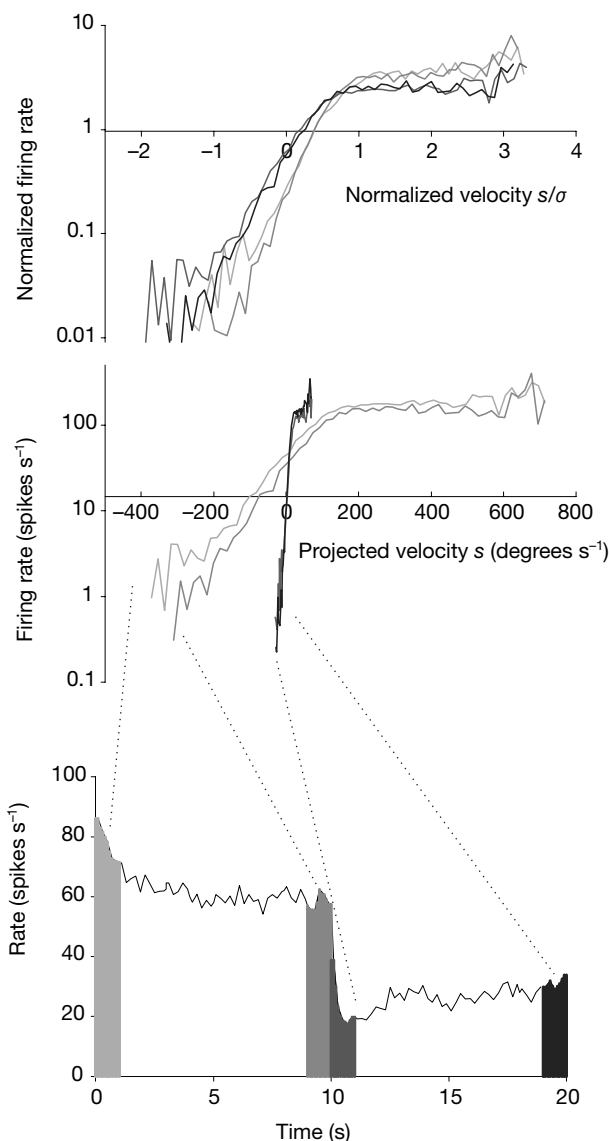


Figure 2 Input/output relations from the switching experiment. Input/output relations were constructed from the 1-s-wide time windows indicated in the rate plot. For the two variance conditions, we compare a window in the initial transient regime after a transition with the final second of the segment, representing the steady state. For a given value, the input/output relation in the transient is identical up to a multiplicative factor to that in the final window, showing that the rescaling of the input by variance is completed in under 1 s. The multiplicative factor stems from the modulation of the rate by the slowly varying adaptive process. The top graph shows the same plots in normalized units. For a given variance, the curves now overlay. The curves for the two different variances are very similar, differing only by a slight shift.

slow, transient change in spike rate after a sudden change in statistics is associated with an evolution of the input/output relation, but only in that the output is multiplied by a factor; the rescaling of the input seems to happen almost immediately on a timescale much shorter than the change in rate. If adaptation involved a single process—a change in neural threshold or the adjustment of inhibitory inputs—then the continuous changes in spike rate during the adaptation transient would be accompanied by a continuous stretching or compression of the input/output relation. Instead the stretching happens rapidly, and scaling of the output is independent of the processes that rescale the inputs. The rate dynamics seem to be a separate degree of freedom in the coding scheme, and may be an independent channel for information transmission.

Rescaling of inputs serves, under stationary conditions, to optimize information transmission¹⁰; it would be attractive if we could observe this optimization dynamically during the adaptation transient. Previous efforts^{16,22} to measure information transmission during adaptation have used the stimulus reconstruction method²³. Here we use a direct method⁶ that estimates the local information transmission rate without any assumptions about the structure of the code and allows us to follow very rapid changes (see Methods). The information transmission as a function of time throughout a switching cycle, with a switch at $t = 0$, is shown in Fig. 3. First, the information per spike is a constant 1.5 bits in the two halves of the experiment, suggesting that maintaining a fixed information per spike may be a design principle of the system²⁴. After a downward switch, the information per spike dips briefly and recovers with a time constant of ~ 40 ms, returning to its maximal level within 100 ms—much faster than the return of the spike rate to steady state. After an upward switch, there is no detectable transient in the

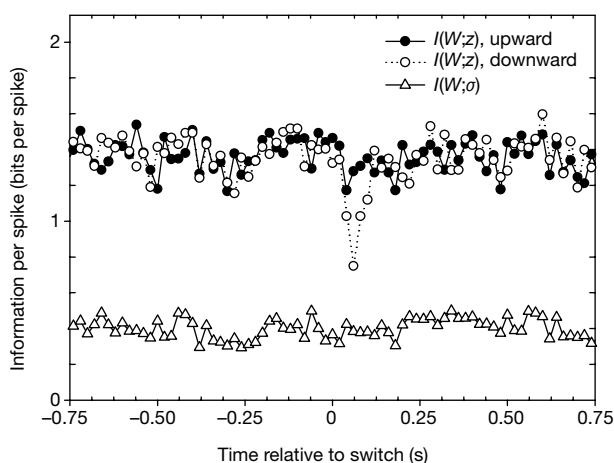


Figure 3 Time dependence of the information transmission. The information/spike was computed using a switching experiment in which the fast component $z(t)$ of the stimulus is a set of 40 fixed, white-noise ‘probe’ segments of 2 s in duration, presented in random order. These were modulated by a square wave envelope for a period of 4 s. The probes were out of phase with the modulation so that each probe straddled a transition from σ_1 to σ_2 . The response of the cell was represented by binary ‘words’ from ten consecutive 2-ms time bins. The presence of a spike in a single bin is represented by 1, the absence by 0. We compute how much information/word the distribution of words at time t (relative to the switch) conveys about which probe was used, and how much information/word is transmitted about the variance σ (see Methods). The information/word is divided by the mean number of spikes per word to obtain the information/spike. The information/spike is plotted as a function of time surrounding an upward and a downward switch in σ . After a downward switch, there is a dip in the information transmission for a duration of <100 ms; after an upward switch, no change in the information transmission rate is detectable. The steady-state information/spike is the same in the two regimes, despite significant differences in spike rate. The information/spike conveyed about σ is also shown—it remains at a steady rate throughout the experiment.

information transmission. Adaptation to variance changes requires that the system gather evidence that signals come from a new probability distribution and are not just outliers in the old distribution. The speed of this sampling is limited by the noise level in the system. As discussed below (see Methods) the minimal source of noise in the photoreceptors sets a physical limit of about 40 ms for the adaptation timescale after a decrease in variance given the conditions of our experiments. Adaptation after an increase in variance can happen more rapidly²⁵. The recovery timescale that we derive theoretically and observe experimentally is remarkably fast, given that the stimulus correlation time is around 10 ms as a result of filtering by the photoreceptors: the system receives barely a few independent samples from the new ensemble. Both the timescale of the information recovery and the asymmetry between increasing and decreasing variance are consistent with adaptation at a speed near the limit set by physical and statistical considerations.

Resolving ambiguity

The fast recovery of information about the rapid variation in the

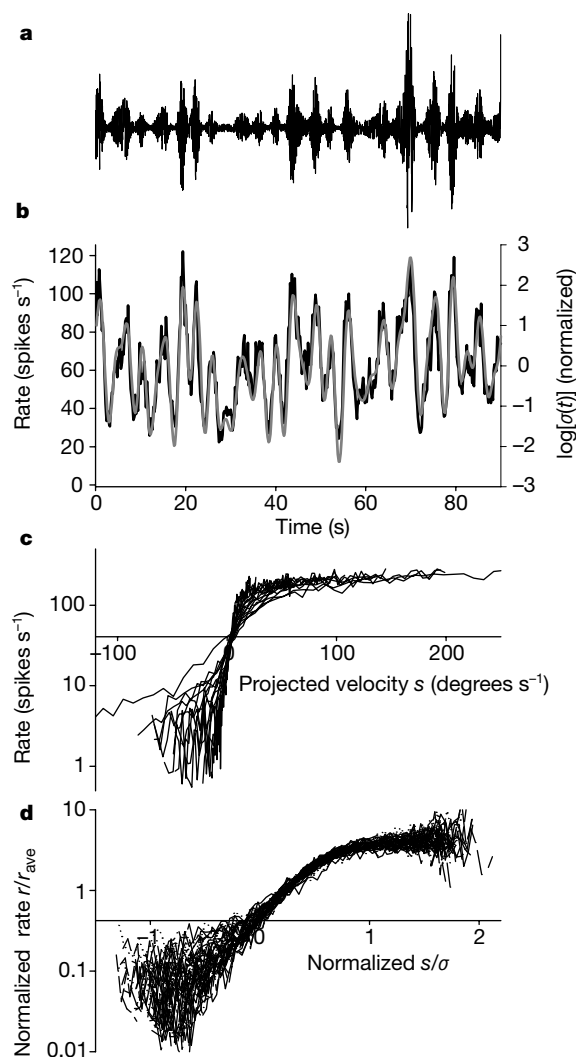


Figure 4 A stimulus with randomly modulated variance. **a**, The stimulus is a white-noise velocity signal modulated by an exponentiated random gaussian process, with correlation time 3 s and cycle length 90 s. **b**, Averaged rate. Overlaid is the logarithm of $\sigma(t)$ for this experiment, normalized to zero mean and unit standard deviation. **c**, Input/output relations from selected windows of width 3 s throughout the experiment. **d**, The same input/output relations as in **c**, rescaled along the x-axis by the local standard deviation, and along the y-axis by the average rate.

stimulus means that one of the possible problems of adaptation is solved, that is, the slow relaxation of the spike rate does not correspond to a similarly long period of inefficient coding. However, the rapid adjustment of the neuron's coding strategy could exacerbate the problem of ambiguity: if the cell encodes a continuously rescaled version of the input signal, how can an observer of the spike train know the overall scale of the stimulus? Although it is possible that other neurons in the network might encode the scale factor, our information-theoretic analysis suggests that H1 itself retains this information; at least in the context of the switching experiment the spike train provides information about the variance at a nearly constant rate per spike (Fig. 3). However, it would be useful if one could extract this information to generate a running estimate of the stimulus variance. An approach to this problem is shown in Fig. 4, where we deliver stimuli in which the variance itself varies at random with a correlation time of 3 s. The result is an intermittent stimulus that is locally gaussian, capturing some of the statistical properties of natural signals in vision²⁶ and in audition²⁷. Under these conditions the spike rate provides an almost linear readout of the logarithm of the variance. At the same time, as before, the system shows continuous adaptation: the input/output relations were calculated in bins throughout the experiment (Fig. 4c), and when the input is normalized by the local standard deviation, the curves coincide (Fig. 4d). This suggests that fine structure of the normalized signal is encoded by the precise placement of spike patterns, generating the information $I(W;z)$ that we measure in Fig. 3, whereas the normalization factor or overall dynamic range of the signal is coded by slower variations in spike rate.

With a sudden change in variance, however, the spike rate remains ambiguous because the representation of the variance is confounded with the time since the last sudden change. Indeed, the stimulus-dependent decay time of the rate could be viewed as coding the time since the switch²⁸. Although the rate may not encode the variance well, this information is present in other statistics of the spike train, as hinted at in ref. 29. Instead of averaging to obtain the time-dependent spike rate, we can collect, for example, all of the interspike intervals, sorted by the order in which they occur after the switch. More than a second after the

switch, when the interval distributions (Fig. 5) have reached a steady state, these steady-state distributions for the two variances are distinct—the mode interval is shifted (Fig. 5); these and other data¹⁰ indicate that the interval distribution provides a 'fingerprint' for the variance, an example of distributional coding³⁰. Figure 5 further shows that by the second or third interval, the distributions are nearly indistinguishable from the steady state. Thus, even while the rate is changing slowly, providing ambiguous information about the stimulus variance, the interspike intervals become typical of the new variance after just a few spikes. We can construct a simple readout scheme for deciding whether the variance is large or small (see Methods), and this readout reaches reliable discrimination with just a few spikes (Fig. 5c), corresponding to a decision time of only 43 ms after a switch to larger variance or 120 ms after a switch to lower variance. These timescales are comparable to those for recovery of information about the normalized stimulus (Fig. 3), and the limit set by the need to gather statistics about the noisy inputs. We conclude that within 100 ms the system both adjusts its coding strategy to optimize information about the rapid (normalized) stimulus variations and provides an unambiguous signal about the overall stimulus scale.

It may seem paradoxical that the spike rate changes over a timescale of seconds when the interspike interval distribution reaches steady state in roughly 100 ms. This can be understood from the structure of the interval distributions, which have three main features: (1) an exponential tail at long intervals; (2) a peak at a preferred interval; (3) a refractory 'hole' at small intervals. A small change in the weight of the tail, or in its slope, can cause significant changes in spike rate, thus the interval distribution can appear almost stationary while the rate changes, and conversely counting spikes (to measure the rate directly) can obscure large changes in the structure of the distribution at shorter times. Both the preferred interval and the shape of the refractory hole depend strongly on the stimulus, in particular on its variance, and these are the features that provide the basis for rapid, reliable discrimination of the variance change (Fig. 5). These effects can be seen even in the response of simple conductance-based models for single neurons responding to injected currents (B. Agüera y Arcas; personal communication) and

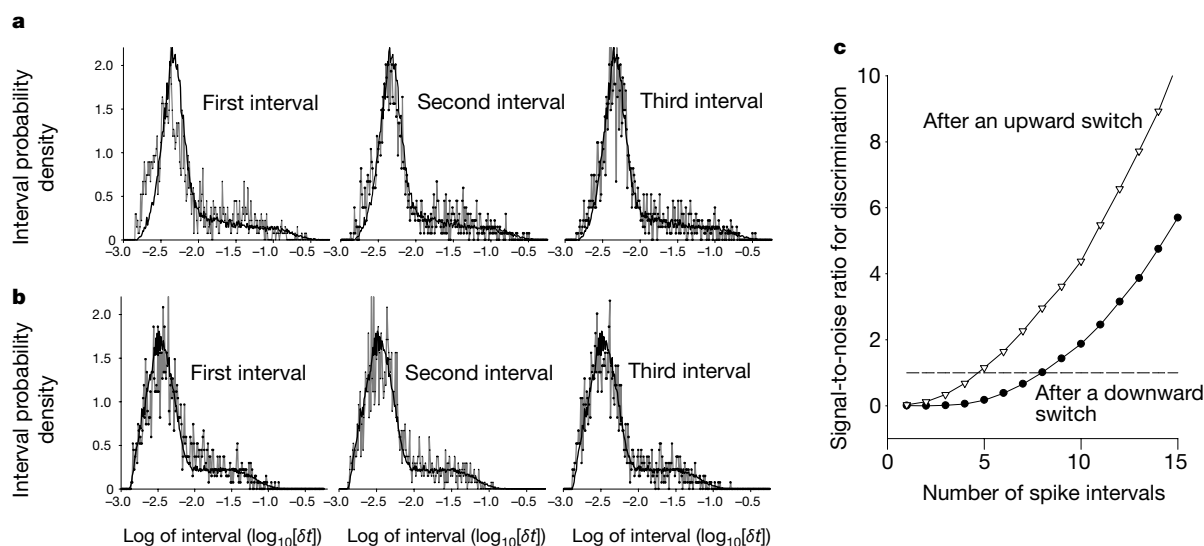


Figure 5 Discrimination using the statistics of interspike intervals in a switching experiment. Intervals were measured for the first several spikes after a switch in the variance, both upward and downward. The intervals shown are all measured in logarithmic units. **a**, The evolution of the interspike-interval probability density is shown for the first, second and third spike intervals after a downward switch, compared with that of all intervals collected after 1 s has elapsed since the switch (the steady-state

distribution, in black). The densities for the individual, sequenced intervals are shown in grey. **b**, After an upward switch, the interval density strongly resembles the steady-state density by the first interval. **c**, The signal-to-noise ratio for discrimination using n spike intervals (see Methods) as a function of the number of intervals after an upward and a downward switch.

in networks of model neurons³¹. This emphasizes that the relative probability of short intervals need not be determined by a fixed absolute or relative refractory mechanism, but can respond to input signals and provide information able to resolve the ambiguities introduced by adaptation.

Discussion

These results suggest a multilayered coding scheme that allows the spike train to convey information about the stimulus through several channels independently: the timing of individual spikes or short spike patterns encodes stimulus features that are normalized to the stimulus ensemble, the statistics of interspike intervals on slightly longer timescales encode the stimulus ensemble, and the spike rate can carry information about the changes in ensemble on yet longer timescales. Although we are not yet able to identify specific mechanisms at work in the visual system of the fly to produce these dynamics, any mechanism must display the existence of multiple timescales, from the <100 ms recovery of information (Fig. 3) to the minute-long relaxation of the firing rate (Fig. 1). Recent biophysical studies³² indicate that such a range (~10³) of timescales can arise even in a single cell from the complex inactivation kinetics of NaII channels. Although multiple timescales could arise also from network dynamics, it is appealing that important aspects of adaptive coding could be implemented by cellular mechanisms. A complementary question concerns the readout of information, for example from the distribution of intervals. This is a classical problem in neural coding, and it has been demonstrated that single neurons are sensitive to the temporal pattern of synaptic inputs³³. Recent work shows how the diverse dynamics of synapses can extract and transmit different aspects of a spike train^{34,35}, providing specific mechanisms for readout of a distributional code.

Many features of adaptation in H1 have analogues in the mammalian visual system. Adaptation to the distribution of image velocities in H1 is most closely analogous to the adaptation to contrast distributions in retinal ganglion cells^{8,36}, where many different gain-control processes span multiple spatial and temporal scales^{37,38}. As with H1, statistical adaptation in retinal ganglion cells is observed at the output of a network; however, recent work shows that aspects of this adaptation can be seen in the response of individual ganglion cells to injected currents³⁹, strengthening the connection to ion channel dynamics. Several nonlinear effects in neural responses in visual cortex have been described in terms of a 'normalization' model⁴⁰ that has a conceptual connection to the rescaling of the input/output relation in H1. Although the phenomena in cortex usually are not described as adaptation to the distribution of inputs, it has been emphasized that the form of these nonlinearities may be matched to the distribution of natural scenes, so that as in H1 their function would be to enhance the efficiency of the neural code⁴¹. The present work shows that such effects can occur very rapidly, and that this rapidity is expected if the system is statistically efficient, so that nearly instantaneous events may nonetheless reflect adaptation to the distribution of inputs.

In traditional demonstrations of adaptation (including in H1 (refs 9, 19)) a long-duration adapting stimulus is presented, followed, after some delay to allow a return to baseline, by a standard probe signal. This adapt-and-probe design has the advantage of showing explicitly that adaptation involves a memory of the adapting stimulus. The difficulty is that it requires a clear separation of timescales between the adaptation process (or its decay) and the transient response to the probe. Instead we find that the nonlinear input/output relation of H1 can change on a timescale comparable to its transient response, and that adaptation involves an apparently continuous range of timescales from tens of milliseconds to minutes. Furthermore, there is evidence for adaptation to the variance and correlation time of specific stimulus features

(motion) as well as to the more general correlation properties of the visual input⁹. Adaptation to statistics over multiple timescales is almost impossible to characterize fully in an adapt-and-probe design, and so here and in ref. 10 we have tried to characterize the response of the neuron to stimuli as they appear in their statistical context. Statistical adaptation can equally well be considered as an example of context dependence in the neural code. According to information theory efficient codes are always context-dependent in this sense, there are limits to how rapidly context can be established, and there are tradeoffs between context-specific coding and ambiguity about the context itself. The neural code for visual motion in H1 seems to be optimal in all these senses, encouraging us to think that information theory may provide a framework for thinking about context dependence of neural coding and computation more generally. □

Methods

Preparation

A fly *C. vicina*, immobilized with wax, viewed a Tektronix 608 display through a circular diaphragm (diameter: 14 horizontal interommatidial spacings), while action potentials were recorded extracellularly. Data in each figure come from a single fly, but all experiments were duplicated on several flies, yielding similar results.

Stimulus

A computer-controlled random, vertical bar pattern, with frames drawn every 2 ms, moves horizontally with velocity $S(t) = \sigma(t)z(t)$, where $z(t)$ is a sequence of unit variance, uniformly distributed random numbers, drawn independently every $\tau_z = 2$ ms. The standard deviation, $\sigma(t)$, varies on a characteristic timescale $\tau_\sigma \gg \tau_z$.

Input/output relation and dimensionality reduction

The input/output relation is the probability distribution $P(\text{spike}|\text{stimulus})$. Accumulating the full distribution is impractical owing to the high dimensionality of the time-dependent stimulus history before each spike. Previous analysis of H1 shows that only a few dimensions control spiking probability. Reverse correlation^{44,2} identifies one of these dimensions: spikes occur at times $\{t_i\}$, and we compute the spike-triggered average stimulus,

$$h(t) = \sum_{i=1}^N S(t - t_i)/N$$

We normalize $h(t)$ to $h_n(t)$ to produce unit gain for a constant velocity. The projections, $s_i = \int dt h_n(\tau) S(t_i - \tau)$, of the stimulus preceding each t_i onto the relevant stimulus dimension h_n , provide samples of the response-conditional distribution²⁶ $P(s|\text{spike})$. The input/output relation follows from Bayes' rule, $P(\text{spike}|s) = P(s|\text{spike})P(\text{spike})/P(s)$, where $P(s)$ is the (gaussian) distribution of all projected stimuli from the experiment and $P(\text{spike})$ is the mean spike rate.

Physical limits to the speed of adaptation

Adaptation to a change in variance of the input signal requires enough time to make a reliable measurement of that variance. This trade between speed and reliability depends on what the system can assume about the nature of the inputs. Suppose we observe a signal $x(t)$ for a time τ and must decide from which of two gaussian distributions it was drawn. For large times τ the probability of error in this decision is dominated by an exponential decay, $P_{\text{error}} \propto \exp(-k\tau)$, with

$$k = \frac{1}{4} \left(\int \frac{d\omega}{2\pi} \left[2 - \frac{X_+(\omega)}{X_-(\omega)} - \frac{X_-(\omega)}{X_+(\omega)} \right] \right)^2 \left(\int \frac{d\omega}{2\pi} \left[1 - \frac{X_+(\omega)}{X_-(\omega)} \right] \right)^{-1}$$

where $X_+(\omega)$ and $X_-(\omega)$ are the power spectra of the two gaussian processes with standard deviations σ_+ and σ_- respectively^{43,44}. Here, $x(t)$ is the brain's internal representation of the angular velocity stimulus, containing components both from the presented signal and from noise. Comparing the contrast signals generated by our visual stimulus with photoreceptor voltage noise measured under comparable conditions⁴⁵ we find that the signal-to-noise ratio is very high up to some critical frequency, $f_c \approx 50$ –100 Hz, above which the effective contrast noise rises steeply. This suggests that we approximate $X_-(\omega)$ as signal-dominated for $\omega < 2\pi f_c$ and noise-dominated for $\omega > 2\pi f_c$. As the two signals are gaussian, their power spectra have the same shape but different variance; this difference is large ($\times 10^2$) under the conditions of Fig. 4.

With these approximations we find $k = f_c/2$, so that the fastest possible relaxation time is about 20–40 ms. Additional noise sources in the visual pathway will necessitate longer times for reliable adaptation. Here the error probability is asymmetric: the prefactor of the exponential is larger for misidentifying a small variance signal as a large variance, predicting that transients will be easier to observe after a decrease in variance. An alternative is to consider continuous estimation of the variance, in which case the asymmetry between increasing and decreasing variance appears as a difference in the timescale of relaxation rather than as a prefactor²⁵.

Information transmission

We calculate information conveyed about the stimulus $S(t) = z(t)\sigma(t)$ as a function of time. At time t , z is one of N probe sequences $\{z_i\}_{i=1,\dots,N}$ and σ one of $\{\sigma_1, \sigma_2\}$. The cell's responses are expressed as binary 'words' (see Fig. 3 legend), discretized at 2 ms, which is approximately the cell's absolute refractory time. Reliable sampling of word distributions requires a compromise between long words and high resolution. Much larger data sets allow finer discretization, possibly yielding higher information rates⁶, but are unlikely to affect our conclusions qualitatively. Interpreted strictly, our results apply only to the case of 2 ms resolution.

With $P_o(W(t)|z)$ —the distribution of word responses at time t for a given probe, z , modulated by σ —the information/word about the probe is

$$I_o(W(t); z) = H[P_o(W(t))] - \sum_{j=1}^N P(z_j) H[P_o(W(t)/z_j)]$$

where H is the entropy of the word distribution:

$$H[P(W(t))] = - \sum_W P(W(t)) \log_2[P(W(t))]$$

Similarly, the information that the distribution of words provides about $\sigma(t)$ is computed from $P_z(W(t)/\sigma)$:

$$I_z(W(t); \sigma) = H[P_z(W(t))] - \sum_{j=1}^2 P(\sigma_j) H[P_z(W(t)/\sigma_j)]$$

We average this quantity over all probes z_i to obtain the average information, $I(W(t); \sigma)$.

Discrimination

We wish to determine whether a sequence of n interspike intervals $(\delta_1, \delta_2, \dots, \delta_n)$ is the response to a stimulus with standard deviation σ_1 or σ_2 . Optimal discrimination requires computing the log-likelihood ratio,

$$D_n = \log[P(\delta_1, \delta_2, \dots, \delta_n | \sigma_1)] / P(\delta_1, \delta_2, \delta_3, \dots, \delta_n | \sigma_2)]$$

but the multidimensional distributions again are difficult to sample. We obtain a lower bound to discrimination performance using a simpler decision variable which would be optimal if there were no significant correlations among successive intervals,

$$D_n = \sum_{i=1}^n \log[P(\delta_i | \sigma_1) / P(\delta_i | \sigma_2)]$$

We measure the reliability of discrimination by the signal-to-noise ratio, $\text{SNR} = \langle D_n \rangle^2 / \langle D_n^2 \rangle - \langle D_n \rangle^2$.

Received 13 March; accepted 5 July 2001.

1. Attneave, F. Some informational aspects of visual perception. *Psych. Rev.* **61**, 183–193 (1954).
2. Barlow, H. B. in *Sensory Communication* (ed. Rosenbluth, W. A.) 217–234 (MIT Press, Cambridge, Massachusetts, 1961).
3. Laughlin, S. B. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch.* **36c**, 910–912 (1981).
4. Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. *Spikes: Exploring the Neural Code* (MIT Press, Cambridge, Massachusetts, 1997).
5. Berry, M. J., Warland, D. K. & Meister, M. The structure and precision of retinal spike trains. *Proc. Natl Acad. Sci. USA* **94**, 5411–5416 (1997).
6. Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. & Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **80**, 197–200 (1998).
7. Reinagel, P. & Reid, C. Temporal coding of visual information in the thalamus. *J. Neurosci.* **20**, 5392–5400 (2000).
8. Smirnakis, S., Berry, M. J., Warland, D., Bialek, W. & Meister, M. Adaptation of retinal processing to image contrast and spatial scale. *Nature* **386**, 67–73 (1997).
9. de Ruyter van Steveninck, R. R., Bialek, W., Potters, M., Carlson, R. H. & Lewen, G. D. in *Natural and Artificial Parallel Computation: Proc. of the Fifth NEC Res. Symp.* (ed. Waltz, D. L.) 21–41 (SIAM, Philadelphia, 1996).
10. Brenner, N., Bialek, W. & de Ruyter van Steveninck, R. Adaptive rescaling maximizes information transmission. *Neuron* **26**, 695–702 (2000).
11. Wainwright, M. Visual adaptation as optimal information transmission. *Vision Res.* **39**, 3960–3974 (1999).
12. Francheschini, N., Riehle, A. & le Nestour, A. in *Facets of Vision* (eds Hardie, R. C. & Stavenga, D. G.) 360–390 (Springer, Berlin, 1989).
13. Hausen, K. in *Photoreception and Vision in Invertebrates* (eds Ali, M.) 523–559 (Plenum, New York, 1984).

14. Schilstra, C. & van Hateren, J. H. Blowfly flight and optic flow. I. Thorax kinematics and flight dynamics. *J. Exp. Biol.* **202**, 1481–1490 (1999).
15. Land, M. F. & Collett, T. S. Chasing behaviour of houseflies (*Fannia canicularis*). *J. Comp. Physiol.* **89**, 331–357 (1974).
16. Clague, H., Theunissen, F. & Miller, J. P. Effects of adaptation on neural coding by primary sensory interneurons in the cricket cercal system. *J. Neurophysiol.* **77**, 207–220 (1997).
17. Fairhall, A. L., Lewen, G., Bialek, W. & de Ruyter van Steveninck, R. R. in *Advances in Neural Information Processing Systems 13* (eds Leen, T. K., Dietterich, T. G. & Tresp, V.) 124–130 (MIT Press, Cambridge, Massachusetts, 2001).
18. Thorson, J. & Biederman-Thorson, M. Distributed relaxation processes in a sensory adaptation. *Science* **183**, 161–172 (1974).
19. de Ruyter van Steveninck, R., Zaagman, W. H. & Mastebroek, H. A. K. Adaptation of transient responses of a movement-sensitive neuron in the visual system of the blowfly *Calliphora erythrocephala*. *Biol. Cybern.* **54**, 223–226 (1986).
20. Borst, A. & Egelhaaf, M. Temporal modulation of luminance adapts time constant of fly movement detectors. *Biol. Cybern.* **56**, 209–215 (1987).
21. van Hateren, J. H. Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation. *J. Comp. Physiol. A* **171**, 157–170 (1992).
22. Warland, D. *Reading Between the Spikes: Real-time Processing in Neural Systems*. Thesis, Univ. California at Berkeley (1991).
23. Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R. & Warland, D. Reading a neural code. *Science* **252**, 1854–1857 (1991).
24. Schneidman, E., Brenner, N., Tishby, N., de Ruyter van Steveninck, R. & Bialek, W. in *Advances in Neural Information Processing Systems 13* (eds Leen, T. K., Dietterich, T. G. & Tresp, V.) 159–165 (MIT Press, Cambridge, Massachusetts, 2001).
25. deWeese, M. & Zador, A. Asymmetric dynamics in optimal variance adaptation. *Neural Comp.* **10**, 1179–1202 (1998).
26. Ruderman, D. L. & Bialek, W. Statistics of natural images: scaling in the woods. *Phys. Rev. Lett.* **73**, 814–817 (1994).
27. Nelken, I., Rotman, Y. & Yosef, O. B. Response of auditory-cortex neurons to structural features of natural sounds. *Nature* **397**, 154–156 (1999).
28. Hopfield, J. J. Transforming neural computations and representing time. *Proc. Natl Acad. Sci. USA* **93**, 15440–15444 (1996).
29. de Ruyter van Steveninck, R. & Bialek, W. Real-time performance of a movement sensitive in the blowfly visual system: information transfer in short spike sequences. *Proc. R. Soc. Lond. Ser. B* **234**, 379–414 (1988).
30. Perkel, D. & Bullock, T. H. Neural coding: a report based on an NRP work session. *Neurosci. Res. Prog. Bull.* **6**, 3 (1968).
31. Wang, Y. & Wang, W. D. Information coding via spontaneous oscillations in neural ensembles. *Phys. Rev. E* **62**, 1063–1068 (2000).
32. Toib, A., Lyakhov, V. & Marom, S. Interaction between duration of activity and time course of recovery from slow inactivation in mammalian brain Na^+ channels. *J. Neurosci.* **18**, 1893–1903 (1998).
33. Segundo, J. P., Moore, G. P., Stensaas, L. J. & Bullock, T. J. Sensitivity of the neurones in *Aplysia* to temporal pattern of arriving impulses. *J. Exp. Biol.* **40**, 643–667 (1963).
34. Markram, H., Gupta, A., Uziel, A., Wang, Y. & Tsodyks, M. Information processing with frequency-dependent synaptic connections. *Neurobiol. Learn. Mem.* **70**, 101–112 (1998).
35. Gerstner, W., Kreiter, A., Markram, H. & Herz, A. Neural codes: firing rates and beyond. *Proc. Natl Acad. Sci. USA* **94**, 12740–12741 (1997).
36. Meister, M. & Berry, M. J. The neural code of the retina. *Neuron* **22**, 435–450 (1999).
37. Shapley, R. M. & Victor, J. D. The contrast gain control of the cat retina. *Vision Res.* **19**, 431–434 (1979).
38. Brown, S. & Masland, R. Spatial scale and cellular substrate of contrast adaptation by retinal ganglion cells. *Nature Neurosci.* **4**, 44–51 (2001).
39. Kim, K. J. & Rieke, F. Temporal contrast adaptation in the input and output signals of salamander retinal ganglion cells. *J. Neurosci.* **21**, 287–299 (2001).
40. Carandini, M., Heeger, D. J. & Movshon, J. A. Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* **17**, 8621–8644 (1997).
41. Simoncelli, E. P. & Schwartz, O. in *Advances in Neural Information Processing Systems 11* (eds Kearns, M. S., Solla, S. A. & Cohn, D. A.) 166–172 (MIT Press, Cambridge, Massachusetts, 1999).
42. Wiener, N. *Extrapolation, Interpolation and Smoothing of Time Series* (MIT Press, Cambridge, Massachusetts, 1949).
43. Feynman, R. P. & Hibbs, A. R. *Path Integrals and Quantum Mechanics* (McGraw Hill, New York, 1965).
44. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966).
45. de Ruyter van Steveninck, R. R. & Laughlin, S. B. The rate of information transfer at graded-potential synapses. *Nature* **379**, 642–645 (1996).

Acknowledgements

We thank B. Agüera y Arcas, T. Adelman and N. Brenner for discussions, and R. Petersen and N. Ulanovsk for comments on the manuscript.

Correspondence and requests for materials should be addressed to A.L.F. (e-mail: adrienne@research.nj.nec.com).

Discovery of three lead-rich stars

S. Van Eck*, S. Goriely*, A. Jorissen* & B. Plez†

* Institut d'Astronomie et d'Astrophysique, Université Libre de Bruxelles, Boulevard du Triomphe, C.P. 226, B-1050 Bruxelles, Belgium

† GRAAL, Université Montpellier II, cc072, F-34095 Montpellier cedex 05, France

About half of the stable nuclei heavier than iron are believed to be synthesized during the late stages of evolution of stars with masses in the range 0.8–8 solar masses. These elements are then expelled into the interstellar medium through stellar winds after being 'dredged up' towards the surface of the stars. These processes occur when the star is in the 'asymptotic giant branch' (AGB) phase of its life. Nuclei (mainly iron) deep inside the star slowly capture neutrons and progressively build up heavier elements (the 's-process'). For AGB stars that formed early in the history of the Galaxy, and that therefore have very low abundances of elements heavier than helium ('metals'), models¹ predict that the s-process will accumulate synthesized material with atomic weights in the Pb–Bi region. Such stars will therefore have large overabundances of lead relative to other heavy elements. Here we report the discovery of large amounts of lead in three metal-poor stars (HD187861, HD196944 and HD224959). Our analysis shows that these stars are more enriched in lead than in any other element heavier than iron. The excellent agreement between the observed and predicted abundances reinforces our current understanding of the detailed operation of the s-process deep in the interiors of AGB stars.

Recent studies^{2–4} have suggested that the transport processes accompanying a deep mining phase, called the third dredge-up, could induce some partial mixing of protons below the convective hydrogen envelope into the helium-burning carbon-rich layers. This partial mixing of protons triggers the chain of reactions $^{12}\text{C}(\text{p},\gamma)^{13}\text{N}(\beta)^{13}\text{C}(\alpha,\text{n})^{16}\text{O}$ that is held responsible for the large neutron irradiation required to synthesize heavy elements by the s-process.

Although this scenario is nowadays the most widely accepted one, its main ingredient (the partial mixing of protons in the deep carbon-rich layers) cannot yet be derived from first physical principles. It is therefore of prime importance to devise predictions that may test this proton-mixing scenario against abundance observations. One such prediction is that low-metallicity AGB stars should exhibit large overabundances of Pb–Bi as compared to lighter s-elements. Such stars would be called 'Pb stars'¹. They are characterized not only by large [Pb/Fe] and [s/Fe] abundance ratios, but also by large [Pb/s] abundance ratios, where

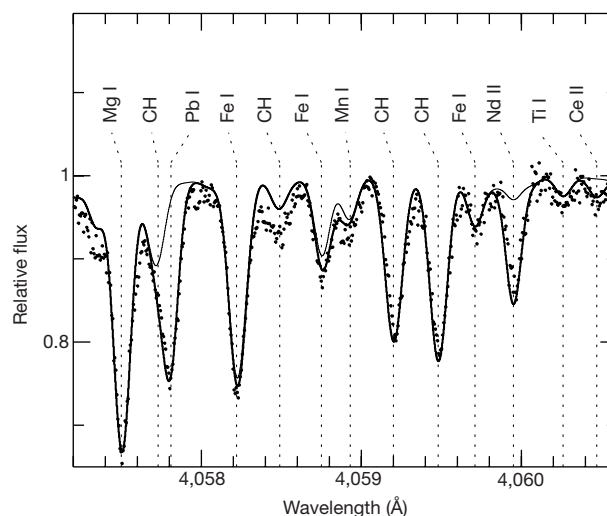


Figure 1 Comparison between observed and synthetic spectra of HD 196944 around the Pb I line at 405.781 nm. The observed spectrum (dots) has been obtained at the ESO 3.6-m telescope and Coudé Echelle Spectrometer delivering a spectral resolution $\lambda/\Delta\lambda$ of 135,000 at 405 nm. Synthetic spectra corresponding to [s/Fe] = 0 (thin solid line; in this case the Pb line is not visible) and [s/Fe] given in Table 1 (thick solid line) are also shown.

$[A/X] \equiv \log(N_A/N_X) - \log(N_A/N_X)_\odot$, s stands for any element produced by the s-process and subscript $\odot$ refers to the Solar System abundances. In particular, [Pb/hs] (where hs denotes the 'heavy' s-process elements such as Ba, La or Ce) ratios as large as 1.5 are predicted¹ in AGB stars with [Fe/H] ≤ -1.3 . Within the framework of the partial mixing of protons into the deep carbon-rich layers, this particular prediction is found to be quite robust¹ with respect to the model parameters (like the abundance profile of the protons in the partially mixed layers, or the extent of the partial mixing zone) and uncertainties (for example, reaction rates). All s-process-enriched AGB stars with metallicities [Fe/H] ≤ -1.3 are thus predicted to be Pb stars, independently of their mass or metallicity (provided the partial mixing of protons takes place).

However, Pb stars remained elusive: none of the low-metallicity stars with large Pb overabundances previously reported in the literature^{5–7} met all the above criteria. Indeed, the Pb abundances^{5,6} in the metal-poor stars HD 115444, HD 126238 and CS 22892-052 are attributed almost exclusively to the r-process, rather than the s-process, because they occur with strong enhancements of r-elements such as Eu (the r-process is an alternative neutron-capture process occurring in high-neutron flux environments like supernova explosions). In the very metal-poor ([Fe/H] = -2.7) carbon-

Table 1 Derived element abundances in three CH stars

HD 187861				HD 224959			HD 196944		
[Fe/H]	−1.65*			−1.7*			−2.45*		
C/O	8*			5*			0.87*		
Element	log(ϵ_x)	[X/Fe]	Error	log(ϵ_x)	[X/Fe]	Error	log(ϵ_x)	[X/Fe]	Error
C	8.80*	1.89*	—	8.50*	1.95*	—	7.43	1.32	0.22
Zr	1.7	0.74	0.3	1.3	0.39	0.3	0.85	0.69	0.15
La	0.9	1.35	0.3	1.05	1.55	0.4	−0.5	0.75	0.21
Ce	1.1	1.14	0.4	1.3	1.39	0.4	0.25	1.09	0.15
Pr	—	—	—	—	—	—	−1.0	0.74	0.15
Nd	1.1	1.28	0.3	1.1	1.33	0.3	−0.05	0.93	0.09
Sm	0.4	1.08	0.4	0.5	1.23	0.3	−0.7	0.78	0.17
Pb	2.8	2.40	0.3	2.9	2.55	0.3	1.75	2.15	0.14

Derived abundances $\log(\epsilon_x) (= \log(N_x/N_H) + 12)$ and [X/Fe] for element X for the CH stars HD 187861, HD 196944 and HD 224959. The listed error represents the total estimated error (as described in Table 2 for HD 196944). The atmospheric parameters were taken from refs 13 and 14 for the three stars, as were the abundances marked with an asterisk. All the other abundances were derived in the present work from spectral synthesis in the 404.5–407.1 nm spectral window using model atmospheres⁸ matching the CNO abundances and the metallicity. Solar oscillator strengths were derived from a synthesis of the solar spectrum, except for the 405.781-nm Pb I line, for which the laboratory value $\log gf = -0.22$ has been adopted¹⁶. Although efforts were made to improve upon existing CN and CH (ref. 17) line lists, the larger errors affecting the abundances in the two carbon-rich stars (HD 187861 and HD 196944) are mainly due to inaccuracies remaining in the CN line list.

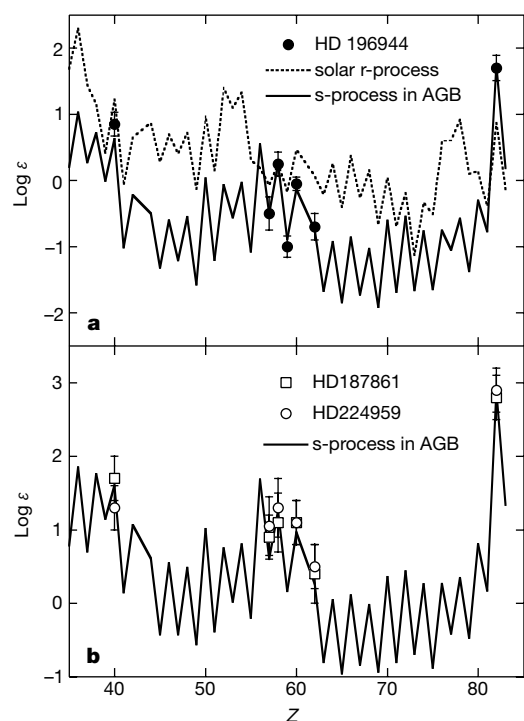


Figure 2 Comparison of the observed abundances with predictions obtained in the framework of the proton-mixing scenario. The abundances observed in HD196944 (panel **a**) and in HD187861 and HD224959 (panel **b**) are compared with the model abundances obtained as described in ref. 1 for stars of the same metallicities as the programme stars. The abundance scale $\log(\epsilon_x)$ stands for $\log(N_x/N_H) + 12$. The error bars on the observations are the root mean square of the uncertainties due (1) to the spectral fit and to the dispersion on the derived abundances when several lines of a given element were used and (2) to the adopted model parameters, that is, the effective temperature, gravity and microturbulence (see Table 2). Panel **a** also shows the solar system r-abundance distribution normalized to Ce. This comparison clearly shows that the large $[Pb/hs] \approx 1.2$ abundance ratio cannot be understood in terms of an r-process enrichment, of which the solar r-process abundances may be considered as a fairly typical representation⁵.

and s-process-rich star LP 625-44, lead was found⁷ to be largely overabundant with respect to iron ($[Pb/Fe] = +2.65$), but not with respect to other s-elements (for example, $[Pb/La] = 0.15$, $[Pb/Ba] = -0.09$), in disagreement with the model predictions.

In the course of a programme of high-resolution spectroscopic observations at the European Southern Observatory 3.6-m telescope, we have discovered three stars that meet all the conditions to be classified as Pb stars.

This observing programme focuses on the study of the nucleosynthesis of elements heavier than iron—with special emphasis on lead—in s-process-rich, low-metallicity AGB stars. It ultimately aims at identifying the origin of Pb–Bi, the heaviest stable elements in the Universe, and at evaluating the respective contributions of the s- and r-processes to their galactic content. This information provides interesting constraints on the operation of these nucleosynthesis processes in stars, and in particular constitutes an important ingredient in the prediction of stellar ages based on the actinide cosmochronometry⁸.

As low-metallicity AGB stars seem to be very rare in the solar neighbourhood nowadays, the family of CH stars represents a good alternative for the purpose of studying the signature of low-metallicity s-processing. These low-metallicity stars are members of binary systems⁹, and their heavy-element overabundances are caused by contamination from ejecta of their companion, formerly an AGB star, presently a dim white dwarf. The surface composition of CH stars thus traces the nucleosynthesis that occurred in a given

Table 2 Effect of uncertainties in model atmosphere parameters on the element abundances derived for HD196944

Element	$\Delta\xi = -0.5 \text{ km s}^{-1}$	$\Delta \log g = +0.3$	$\Delta T_{\text{eff}} = +100 \text{ K}$	σ
C	0	-0.1	+0.2	0.03
Zr	0	+0.1	+0.05	(0.1)
La	+0.05	+0.05	+0.2	(0.1)
Ce	0	+0.1	+0.05	0.1
Pr	0	+0.1	+0.1	0.05
Nd	0	+0.05	+0.05	0.05
Sm	0	+0.1	+0.1	0.1
Pb	+0.03	0	+0.1	(0.1)

Parameters are effective temperature T_{eff} , gravity g and microturbulence ξ . The values adopted for the reference model are taken from ref. 13 ($T_{\text{eff}} = 5,250 \text{ K}$, $\log g = 1.4$, $\xi = 2.0 \text{ km s}^{-1}$). The last column corresponds to the standard deviation of the abundances when derived from different lines. If only one line is available, an error of 0.1 dex (enclosed in parentheses) is assumed. Because the oscillator strength of the Pb I line at 405.781 nm cannot be astrophysically assessed—this line is barely detectable in the Sun—the impact of a $\Delta \log gf = \pm 0.05$ error on the $\log gf$ value (resulting in $\Delta \log \epsilon_{\text{Pb}} = \pm 0.05$) has been included in the errors listed in Table 1.

low-metallicity AGB star. Comparison with model predictions is therefore straightforward, which would not be the case if we were to use surface abundances that would be derived from low-metallicity unevolved field stars (their abundances reflect, on the contrary, the plethora of nucleosynthetic events that shaped the composition of the interstellar medium from which the stars formed).

All three stars observed present large Pb abundances (Fig. 1 and Table 1) and may be considered to be genuine s-process Pb stars, because (1) these s-process-rich stars do exhibit overabundances of Pb with respect to all the other s-process elements, and (2) the Pb overabundance cannot originate from the competing r-process (that hypothesis would indeed lead to an abundance pattern totally incompatible with the observed one; see Fig. 2). The estimated errors on our abundance determinations do not endanger this conclusion (Tables 1 and 2). High spectral resolution is essential to resolve the 405.781-nm Pb line from the nearby CH line at 405.77 nm, and therefore to derive correct lead abundances (see Fig. 1).

The observed surface abundances are in remarkable agreement with the predictions¹ obtained for partial mixing operating in stars of the corresponding metallicities (Fig. 2). As indicated above, there is actually little room for variations at these low metallicities within the framework of the ‘proton-mixing’ scenario (in that sense, the LP 625-44 abundance pattern⁷ appears puzzling, and may call for an alternative scenario¹⁰ operating in extremely low-metallicity stars, provided LP 625-44 is a representative case).

The similarity between the abundance patterns of our three stars is remarkable, and strongly suggests that this pattern constitutes the standard signature of s-processing at low metallicity, in agreement with the model predictions described above. If so, the discovery of the three Pb stars reported here validates the ‘proton-mixing’ scenario for the detailed operation of the s-process in AGB stars. This model is currently the only one capable of explaining such high Pb overabundances in s-process-enriched stars (see Fig. 2). As a consequence, the solar s-process distribution should no longer be thought to result from successive neutron irradiations within a single star¹¹, but rather as a result of the chemical evolution of the Galaxy shaping the s-process abundance distribution¹².

Moreover, to calculate stellar ages cosmochronologically on the basis of actinide abundances we need to separate the contributions of the s- and r-processes to the galactic Pb production. The discovery of low-metallicity s-rich Pb stars makes it possible to constrain the r-process contribution, which controls our prediction of the production of the long-lived actinides. More accurate ages for the oldest stars in our Galaxy using the Th cosmochronometry may thus be obtained⁸. □

Received 13 March; accepted 6 July 2001.

- Goriely, S. & Mowlavi, N. Neutron-capture nucleosynthesis in AGB stars. *Astron. Astrophys.* **362**, 599–614 (2000).
- Gallino, R. *et al.* Evolution and nucleosynthesis in low-mass asymptotic giant branch stars. II. Neutron capture and the s-process. *Astrophys. J.* **497**, 388–403 (1998).

3. Herwig, F., Blocker, T., Schönberner, D. & El Eid, M. Stellar evolution of low and intermediate-mass stars. IV. Hydrodynamically-based overshoot and nucleosynthesis in AGB stars. *Astron. Astrophys.* **324**, L81–L84 (1997).
4. Langer, N., Heger, A., Wellstein, S. & Herwig, F. Mixing and nucleosynthesis in rotating TP-AGB stars. *Astron. Astrophys.* **346**, L37–L40 (1999).
5. Sneden, C., Cowan, J. J., Burris, D. L. & Truran, J. W. Hubble Space Telescope observations of neutron-capture elements in very metal-poor stars. *Astrophys. J.* **495**, 235–245 (1998).
6. Sneden, C. *et al.* Evidence of multiple r-process sites in the early Galaxy: New observations of CS 22892-052. *Astrophys. J.* **533**, L139–L142 (2000).
7. Aoki, W., Norris, J. E., Ryan, S. G., Beers, T. C. & Ando, H. Detection of lead in the carbon-rich, very metal-poor star LP 625-44: A strong constraint on s-process nucleosynthesis at low metallicity. *Astrophys. J.* **536**, L97–L100 (2000).
8. Goriely, S. & Clerbaux, B. Uncertainties in the Th cosmochronometry. *Astron. Astrophys.* **346**, 798–804 (1999).
9. McClure, R. D. & Woodsworth, A. W. The binary nature of the barium and CH stars. III. Orbital parameters. *Astrophys. J.* **352**, 709–723 (1990).
10. Fujimoto, M. Y., Ikeda, Y. & Iben, L. Jr. The origin of extremely metal-poor carbon stars and the search for population III. *Astrophys. J.* **529**, L25–L28 (2000).
11. Käppeler, F., Beer, H. & Wisshak, K. s-Process nucleosynthesis—nuclear physics and the classical model. *Rep. Prog. Phys.* **52**, 945–1013 (1989).
12. Travaglio, C., Galli, D., Gallino, R., Busso, M., Ferrini, F. & Straniero, O. Galactic chemical evolution of heavy elements: From barium to europium. *Astrophys. J.* **521**, 691–702 (1999).
13. Zács, L., Nissen, P. E. & Schuster, W. J. The chemical composition of HD 196944: a carbon and s-process rich, very metal-poor star. *Astron. Astrophys.* **337**, 216–222 (1998).
14. Vanture, A. D. The CH stars. III. Heavy element abundances. *Astron. J.* **104**, 1997–2004 (1992).
15. Alvarez, R. & Plez, B. Near-infrared narrow-band photometry of M-giant and Mira stars: models meet observations. *Astron. Astrophys.* **330**, 1109–1119 (1998).
16. Biémont, E., Garnir, H. P., Palmeri, P., Li, Z. S. & Svanberg, S. New f-values in neutral lead obtained by time-resolved laser spectroscopy, and astrophysical applications. *Mon. Not. R. Astron. Soc.* **312**, 116–122 (2000).
17. Luque, J. & Crosley, D. R. LIFBASE: database and spectral simulation program. SRI International Report MP 99-009 (1999) available at <http://www.sri.com/psd/lifbase>.

Acknowledgements

We thank R. Alvarez, M. Arnould and V. Hill for help and advice. S.G. and A.J. are FNRS Research Associates. S.V. is an FNRS Post-doctoral Researcher. This work is based on observations carried out at the European Southern Observatory in Chile.

Correspondence should be addressed to S.V. (e-mail: svaneck@astro.ulb.ac.be).

Noble-gas-rich chondrules in an enstatite meteorite

Ryuji Okazaki[†], Nobuo Takaoka^{*}, Keisuke Nagao[‡], Minoru Sekiya^{*} & Tomoki Nakamura^{*}

^{*} Department of Earth and Planetary Sciences, Faculty of Sciences, Kyushu University 33, Hakozaki, Fukuoka 812-8581, Japan

[‡] Laboratory for Earthquake Chemistry, Graduate School of Science, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Chondrules are silicate spherules that are found in abundance in the most primitive class of meteorites, the chondrites. Chondrules are believed to have formed by rapid cooling of silicate melt early in the history of the Solar System¹, and their properties should reflect the composition of (and physical conditions in) the solar nebula at the time when the Sun and planets were forming. It is usually believed that chondrules lost all their noble gases at the time of melting^{2–4}. Here we report the discovery of significant amounts of trapped noble gases in chondrules in the enstatite chondrite Yamato-791790, which consists of highly reduced minerals. The elemental ratios $^{36}\text{Ar}/^{132}\text{Xe}$ and $^{84}\text{Kr}/^{132}\text{Xe}$ are similar to those of 'subsolar' gas^{5,6}, which has the highest $^{36}\text{Ar}/^{132}\text{Xe}$ ratio after that of solar-type noble gases⁷. The most plausible explanation for the high noble-gas concentration and

the characteristic elemental ratios is that solar gases were implanted into the chondrule precursor material, followed by incomplete loss of the implanted gases through diffusion over time.

In the course of a study of noble gases in 12 enstatite chondrites (ECs; enstatite is magnesium silicate), we analysed whole-rock samples of Yamato-791790 by total melting, and by mechanical crushing followed by stepped heating. We found ratios of $^{36}\text{Ar}/^{132}\text{Xe}$ (ref. 8) that were up to 50 times higher (Fig. 1 and Table 1) than those of 'Q-gas' trapped in 'phase-Q'⁹, which is the main carrier of noble gases in carbonaceous and ordinary chondrites; $^{36}\text{Ar}/^{132}\text{Xe}$ ratios for 'Q-gas'¹⁰ are as low as 60. The ratios $^{36}\text{Ar}/^{132}\text{Xe}$ and $^{84}\text{Kr}/^{132}\text{Xe}$ for Yamato-791790 are indicative of a high concentration of subsolar gas⁵ (or Ar-rich gas)⁶; this gas was observed in ECs almost 20 years ago, but its trapping phase has not been clear until now. Some ECs with the subsolar signature have Xe isotope ratios that bear a close resemblance to solar Xe⁵. Both the Xe isotope ratios and the high $^{36}\text{Ar}/^{132}\text{Xe}$ ratio suggest a contribution of solar-type noble gas to ECs. However, ECs rich in subsolar gas⁵ are depleted in ^4He and ^{20}Ne , in contrast to the high concentrations of He and Ne found in solar-gas-rich meteorites⁷. Identification of a trapping phase would help to elucidate the origin of subsolar gas, but such a phase has not been clearly identified—although enstatite has been proposed as the carrier of subsolar gas in ECs^{5,6,11}.

In order to investigate the trapping site of subsolar gas in Yamato-791790, we performed a laser-microprobe noble-gas analysis. We prepared five polished thin sections, each section being about 50 μm in thickness. Optical microscopic observation revealed that Yamato-791790 contains chondrules (100–500 μm in diameter), chondrule fragments, and opaque grains (mostly Fe-Ni metal and troilite) with relatively small amounts of matrix materials. Porphyritic or euhedral pyroxene is dominant in most chondrules. We determined the major-element composition with an electron microprobe analyser. All pyroxene grains analysed in this study are low-CaO (< 1 wt%). Low-FeO (< 1 wt%) pyroxene is common in chondrules and matrices, but high-FeO (> 10 wt%) pyroxenes are rare. These petrologic and mineralogical features are consistent with previous reports that Yamato-791790 is an EC of petrologic type 4 showing the imprint of moderate thermal effects, and that it experienced heavier impact events¹² than other ECs. Both sides of the thin sections were examined by optical microscopy to confirm that a portion selected for laser ablation consists of a single component.

Noble gases were extracted by ablating individual components with a pulsed Nd:YAG laser beam (1.064 μm wavelength, 40–70 μm in diameter), and were measured with a modified VG5400 MS-II system at the University of Tokyo¹³. Ablated mass is typically 0.7 μg , ranging from 0.24 to 3.7 μg . Laser extraction efficiencies relative to a conventional melting method were determined to be 1.39, 1.09 and 1.08 for ^{36}Ar , ^{84}Kr and ^{132}Xe , respectively. The concentrations of noble gases listed in Table 1 are corrected for the blank and the extraction efficiencies. Because of the small mass ablated by the laser and the low concentrations of ^{84}Kr and ^{132}Xe in this meteorite, the amount of gas available for the measurement is extremely low, and in some cases, is indistinguishable from the blank. Consequently, errors for ^{84}Kr and ^{132}Xe concentrations became large compared with those for ^{36}Ar . Hence the elemental ratios for the laser ablation analysis in Table 1 and Fig. 1 are corrected only for the extraction efficiencies without the blank correction.

We analysed two large opaque grains, eight matrix portions, and twenty individual chondrules (Table 1). Most of the analysed chondrules are 100–200 μm in diameter, while some (5Ch, 7Ch, 9Ch and 20Ch) are somewhat larger (350, 300, 500–800 and 500 μm , respectively). No laser spots released noticeable amounts of He and Ne, whereas heavier noble gases were detected in most cases. The detected ^{36}Ar was mostly that which had been trapped

[†] Present address: Laboratory for Earthquake Chemistry, Graduate School of Science, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

within this chondrite, because the cosmic-ray exposure age for Yamato-791790 is short (3.3 Myr)¹³, and consequently the concentration of cosmogenic ³⁶Ar was negligible. We found very high ³⁶Ar concentrations, up to $7 \times 10^{-6} \text{ cm}^3 \text{ STP g}^{-1}$, in several chondrules (for example, 5Ch-d, 9Ch-d, 17Ch and 20Ch-a in Table 1). Two examples are shown in Fig. 2. Spots 5Ch-b, 5Ch-c and 5Ch-d in the chondrule 5Ch released much larger amounts of ³⁶Ar than spots 5Mx-a and 5Mx-e in the matrix portions (Table 1 and Fig. 2). The high ³⁶Ar concentrations in the chondrule portions certainly reflect trapped Ar indigenous to the chondrule because the actual spatial resolution of the laser system is about 50–80 μm (1.2 times the pit diameter). Such a notable difference in ³⁶Ar concentrations between the chondrule and the surrounding matrix is also seen in another chondrule, 7Ch (7Ch-a, 7Ch-b, 7Ch-c and 7Ch-e versus 7Mx-d and 7Mx-f). Concentrations of ³⁶Ar in the chondrules 5Ch and 7Ch are uniform within each chondrule, but the centres of other chondrules seem to be more enriched in ³⁶Ar—for example, the centre portion

(20Ch-a) of chondrule 20Ch contains more abundant ³⁶Ar than do the other portions (20Ch-b, 20Ch-c and 20Ch-d). The largest chondrule 9Ch (500–800 μm) analysed in the thin sections also shows higher ³⁶Ar concentrations (Fig. 2) in the centre portions (9Ch-a, 9Ch-c and 9Ch-d) compared to the edge (9Ch-b).

The scatter of ³⁶Ar concentrations is greater between chondrules than within a chondrule (Table 1), but we observed no clear correlation between ³⁶Ar and chondrule size. However, there is a trend that chondrules consisting of porphyritic pyroxene (PP) are rich in trapped Ar, while non-porphyritic-pyroxene (NPP) chondrules are depleted in it (Table 1). This suggests that the abundances of trapped noble gases are related to the chondrule texture—and the texture depends on the maximum temperature and cooling rates during chondrule formation¹⁴.

In contrast to chondrules, ³⁶Ar concentrations for large opaque grains and matrices are very low ($<1 \times 10^{-6} \text{ cm}^3 \text{ STP g}^{-1}$), and elemental ratios lie on a mixing line between 'Q-gas' and the

Table 1 Heavy noble gases in the Yamato-791790 enstatite chondrite

Pit	Component	Mass§ (μg)	³⁶ Ar ($10^{-6} \text{ cm}^3 \text{ STP g}^{-1}$)	⁸⁴ Kr ($10^{-9} \text{ cm}^3 \text{ STP g}^{-1}$)	¹³² Xe ($10^{-9} \text{ cm}^3 \text{ STP g}^{-1}$)	³⁶ Ar/ ¹³² Xe¶	⁸⁴ Kr/ ¹³² Xe¶
BULK1*	WR	11000	1.48	3.33	0.814	1,818	4.09
BULK2*	WR	1770	2.28	4.80	1.08	2,111	4.44
S6P*	WR	344	1.54	3.48	0.799	1,927	4.36
S8P*	WR	2170	1.53	3.97	1.20	1,275	3.31
S6L†	WR	32	1.86	3.38	0.846	2,199	4.00
S8L†	WR	26	2.41	4.81	1.32	1,826	3.64
1Op	OP	3.7	0.53 ± 0.83	12.3 ± 7.2	1.3 ± 3.2	258	3.10
2Op	OP	0.36	0.05 ± 0.07	1.68 ± 0.90	0.43 ± 0.28	176	3.21
3Mx	MX	0.42	0.69 ± 0.73	7.0 ± 5.6	3.0 ± 2.8	228	1.99
4Mx	MX	0.61	0.25 ± 0.47	3.3 ± 3.7	9.1 ± 2.7	84.7	0.764
5Mx-a	MX	1.1	0.27 ± 0.26	7.2 ± 3.4	0.12 ± 0.93	353	6.16
5Ch-b	PP	0.71	4.6 ± 1.4	17.1 ± 6.2	4.5 ± 1.7	723	3.35
5Ch-c	PP	0.71	5.4 ± 1.7	18.6 ± 6.6	3.4 ± 1.6	976	4.20
5Ch-d	PP	0.71	7.1 ± 2.2	27.5 ± 8.5	6.0 ± 1.9	884	3.96
5Mx-e	MX	0.94	0.55 ± 0.33	11.3 ± 4.4	0.7 ± 1.1	376	6.14
6Mx-a	MX	0.70	0.33 ± 0.38	1.7 ± 3.9	0.3 ± 1.4	308	2.85
6Ch-b	PP	0.71	0.79 ± 0.45	2.9 ± 3.9	0.2 ± 1.4	487	3.45
6Ch-c	PP	0.71	2.00 ± 0.74	2.5 ± 3.9	ND	978	3.59
7Ch-a	PP	0.71	2.69 ± 0.92	6.0 ± 4.3	1.0 ± 1.4	880	3.49
7Ch-b	PP	0.71	2.28 ± 0.81	6.3 ± 4.3	1.2 ± 1.4	743	3.42
7Ch-c	PP	0.71	2.08 ± 0.76	5.9 ± 4.3	1.2 ± 1.4	691	3.34
7Mx-d	MX	1.1	0.62 ± 0.32	2.8 ± 2.7	2.6 ± 1.1	235	1.72
7Ch-e	PP	0.71	2.21 ± 0.79	5.3 ± 4.2	6.4 ± 2.0	311	1.36
7Mx-f	MX	1.1	0.17 ± 0.24	1.6 ± 2.6	3.0 ± 1.1	123	1.34
8Mx	MX	0.47	0.07 ± 0.51	20.4 ± 8.3	2.0 ± 2.1	166	5.10
9Ch-a	PP	0.42	3.1 ± 1.2	8.1 ± 5.7	0.7 ± 2.7	720	2.94
9Ch-b	PP	0.85	1.39 ± 0.58	4.2 ± 2.9	1.4 ± 1.4	487	2.19
9Ch-c	PP	0.59	4.9 ± 1.5	10.8 ± 3.9	0.9 ± 1.6	1,356	3.74
9Ch-d	PP	0.88	5.3 ± 1.6	11.6 ± 3.6	1.2 ± 1.1	1,723	4.42
10Ch	PP	0.37	0.29 ± 0.77	5.2 ± 6.1	ND	293	2.80
11Ch	PP	0.64	4.7 ± 1.5	12.0 ± 4.9	7.5 ± 2.4	500	1.65
12Ch	PP	0.29	2.2 ± 1.3	8.9 ± 7.9	2.6 ± 3.9	384	2.19
13Ch	PP	0.42	4.6 ± 1.6	15.9 ± 7.0	1.8 ± 2.7	815	3.56
14Ch	PP	0.42	2.7 ± 1.1	9.8 ± 5.9	0.6 ± 2.7	660	3.27
15Ch	PP	0.32	2.6 ± 1.3	10.5 ± 7.6	1.3 ± 3.6	506	2.76
16Ch-a	PP	0.52	0.33 ± 0.56	6.2 ± 4.6	4.2 ± 2.4	153	1.62
16Ch-b	PP	0.52	0.25 ± 0.56	3.7 ± 4.3	2.6 ± 2.3	179	1.66
17Ch	PP	0.92	5.0 ± 1.5	10.9 ± 3.9	3.3 ± 1.4	967	2.66
18Ch	PP	0.29	1.3 ± 1.0	1.6 ± 9.0	ND	463	3.08
19Ch	PP	0.47	2.4 ± 1.0	15.1 ± 7.4	2.3 ± 2.1	518	4.02
20Ch-a	PP	0.38	5.5 ± 1.8	3.8 ± 3.9	ND	2,231	3.71
20Ch-b	PP	0.31	1.66 ± 0.88	6.1 ± 5.0	ND	738	3.83
20Ch-c	PP	0.44	3.1 ± 1.1	7.6 ± 4.0	ND	1,705	5.90
20Ch-d	PP	0.37	2.8 ± 1.1	5.9 ± 4.3	ND	909	3.11
21Ch	PP	0.94	1.88 ± 0.66	10.7 ± 4.3	1.6 ± 1.1	649	4.44
22Ch	PP	0.71	2.58 ± 0.89	7.0 ± 4.4	1.4 ± 1.4	774	3.42
23Ch	NPP	0.24	ND	6.3 ± 9.1	0.5 ± 4.7	208	2.33
24Ch	NPP	0.47	0.15 ± 0.52	4.9 ± 6.0	0.5 ± 2.1	235	3.38
25Ch	NPP	0.71	ND	4.9 ± 4.2	1.5 ± 1.4	132	2.85

* Noble gases for the whole rock sample extracted by total melting. Uncertainties for concentrations and elemental ratios are estimated to be 10%.

† Noble gases released by laser ablation of sample slices. Concentrations of the sample S6L and S8L are calculated from noble gas abundance corrected for the blank and fused mass without correction for extraction efficiencies.

‡ Abbreviations: WR, whole rock; OP, opaque grain; MX, matrix; PP, porphyritic pyroxene chondrule; NPP, non-porphyritic pyroxene chondrule.

§ Laser-fused mass, estimated from apparent diameter of laser pit, sample thickness, and density calculated from modal abundance of minerals in the analysed area.

|| Concentrations of noble gases released by laser ablation, calculated from laser-released noble gases (for example, Ar-L), blank, extraction efficiency (Eff.), and fused mass; thus

[Ar-conc.] = [(Ar-L/1.39 – blank)/mass] where 1.39 is Eff. for Ar. Errors for concentrations are calculated from propagation of errors for the blank (50%), fused mass (20%), sensitivities (5%), and Eff. (20, 15 and 5% for ³⁶Ar, ⁸⁴Kr, and ¹³²Xe, respectively). The 20% errors for fused mass are estimated from errors for thickness (10%), ablated surface area (15%), and density (10%). Absolute amounts of blank gases measured during the laser-microprobe analysis are ($\text{cm}^3 \text{ STP}$) ³⁶Ar = 5×10^{-13} , ⁸⁴Kr = 5×10^{-15} and ¹³²Xe = 2×10^{-15} with 50% uncertainties, while elemental ratios of the blank gases are ³⁶Ar/¹³²Xe = 245 and ⁸⁴Kr/¹³²Xe = 2.3 with 30% uncertainties. ND, not determined.

¶ Elemental ratios of laser ablation experiments are not corrected for the blank. For example, ³⁶Ar/¹³²Xe = [Ar-L/1.39]/[Xe-L/1.08], where 1.39 and 1.08 are the extraction efficiencies for Ar and Xe, respectively. Uncertainties for ³⁶Ar/¹³²Xe and ⁸⁴Kr/¹³²Xe are estimated to be about 15% from uncertainties for sensitivities (5%) and errors for Eff. ratios between elements (about 15%).

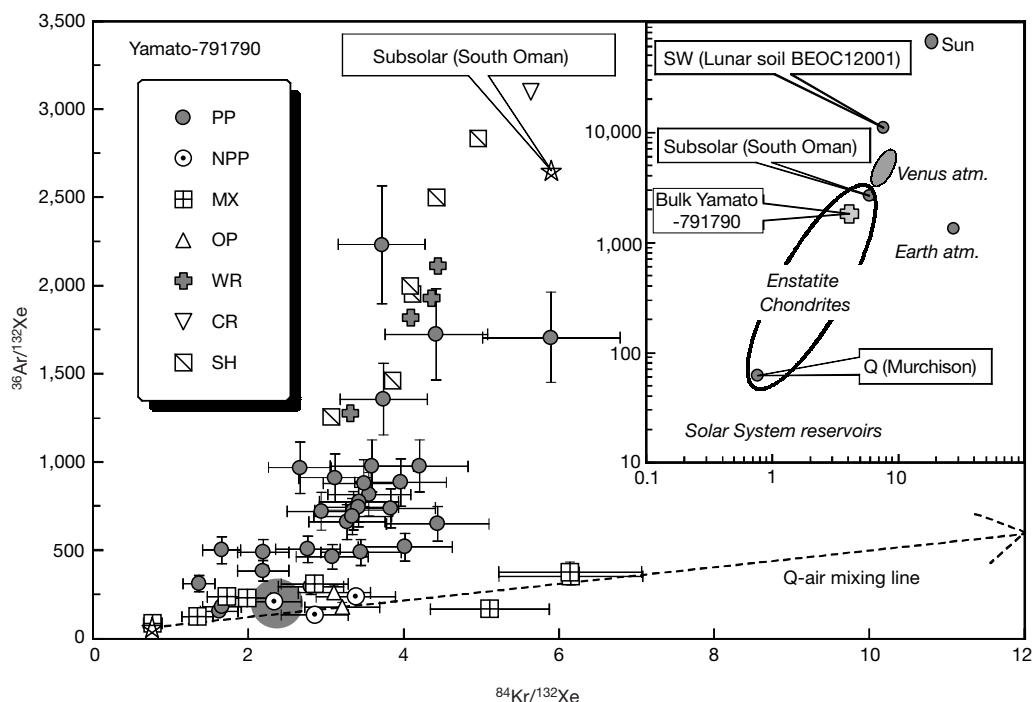


Figure 1 Abundance ratios of noble gases. Inset, noble gases in ECs can be explained by mixing of 'Q-gas' and subsolar gases. Main figure, chondrules of Yamato-791790 are enriched in ^{36}Ar compared to matrices or opaque minerals in this meteorite. Elemental ratios for gases obtained by crushing and step-heating of Yamato-791790, as well as the south Oman EC and 'Q-gas', are plotted for comparison. Elemental ratios of all laser-

released gases are affected by those of the blank represented as a shaded circle at the lower left. Data are shown for the Sun²¹, lunar soil⁷, terrestrial atmosphere²², venusian atmosphere²³, 'Q-gas'¹⁰, the south Oman EC⁵, and crushing and step-heating of Yamato-791790 (ref. 8). Abbreviations: CR, crushing method; SH, step-heating method (1,000–1,600 °C fractions). The other abbreviations are as used in Table 1.

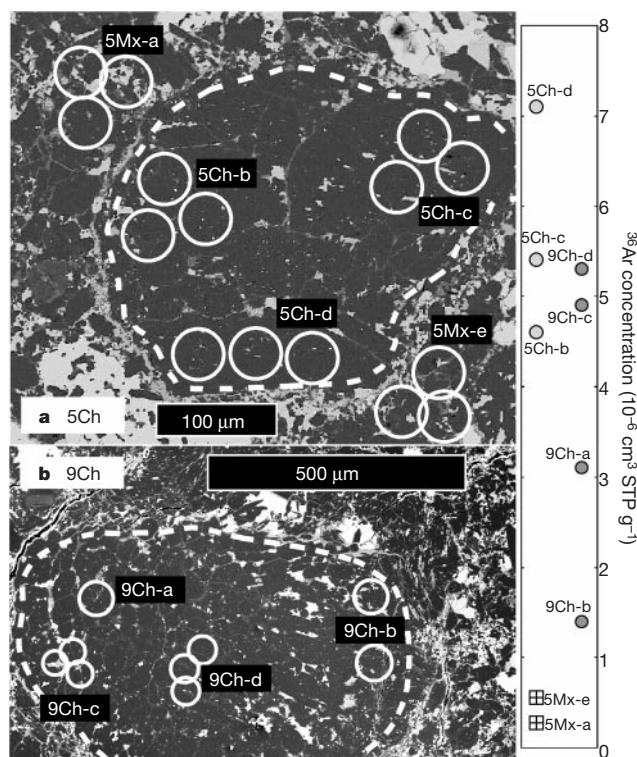


Figure 2 Microphotographs of porphyritic pyroxene chondrules, showing sampling sites, and measured ^{36}Ar concentrations. At the sampling sites, small portions of these chondrules and surrounding matrices are individually extracted by laser ablation, and the released gases measured. **a**, Each analysed portion (consisting of three pits; white circles) is a similar distance from the boundary between the chondrule (5Ch) and matrices (5Mx).

Spots in the chondrule released larger amounts of ^{36}Ar than those in the matrix, which indicates distinctly that the detected ^{36}Ar originates from the interior of the chondrule. **b**, The centre of the chondrule 9Ch shows higher ^{36}Ar concentrations than the marginal region.

terrestrial atmosphere (Fig. 1), suggesting that opaque grains and matrix portions are poor in subsolar gas. On the other hand, one matrix portion (4Mx in Table 1) shows a high ^{132}Xe concentration along with elemental ratios essentially identical to those of 'Q-gas'—this indicates that residual 'phase-Q' is located in the matrix portion, as in other chondrite classes².

How were large amounts of noble gases trapped in chondrules and retained within the chondrule-forming minerals? Implantation of high-energy particles from the young Sun would explain the high ^{36}Ar concentration in silicate materials, as in the case of lunar soils abundant in solar gases⁷. For example, the 'X-wind' model¹⁵ proposed particle irradiation in the early Solar System. Our estimate predicts that diffusive loss¹⁶ of solar-type noble gases from silicate melt can explain $^{20}\text{Ne}/^{36}\text{Ar}/^{132}\text{Xe}$ ratios and abundances of subsolar gas. In the calculation, we assumed that Ne, Ar and Xe of initial concentrations equal to lunar soil⁷ migrated through chondrule-sized materials heated at 1,600 K for 1,000 s. Even the energetic solar flare particles (up to 100 MeV per nucleon) can penetrate at most 1 mm into silicate matter¹⁷, and hence noble-gas implantation would have occurred on fine grains of chondrule-precursor materials. If this is the case for the subsolar gases in chondrules, it is reasonable to expect that chondrules in other chondrite classes might also contain solar-type noble gas—in contrast to the limited reports that most chondrules in other chondrites are free of primordial noble gases^{2–4}. Alternatively, chondrules in each chondrite class might have been produced from different precursor materials, and/or through different heating events during which solar gases trapped in the chondrule precursors would be lost due to high peak temperatures and a slow cooling rate.

On the other hand, astronomical observations have established that young solar-mass stars go through a phase of increased activity (the T Tauri phase), during which particle fluxes are considerably greater¹⁸ than present solar fluxes. If abnormally high-energy particles were available during this active phase, they could have penetrated into the interior of chondrules; thus noble-gas implantation onto the surface of the EC parent body at the inner region of protosolar nebula could be responsible for the subsolar gas in chondrules. The elemental ratios of subsolar gas in chondrules can be explained by diffusive loss of solar-type noble gases, as mentioned above, while the subsolar-gas depletion in matrix portions could be the result of diffusion loss through small silicate grains (5 μm across). However, explaining the absence of noble gases in large opaque grains remains difficult.

Other mechanisms for trapping noble gases are adsorption on chondrule precursor materials, and solution into enstatite melt. The high concentrations of noble gases found in chondrules means that a high adsorption efficiency would have been required—this, in turn, requires very low temperatures. And such low-temperature adsorption would have brought a large amount of water into chondrule precursor materials; but there are no hydrous minerals in ECs¹⁹, so we consider this mechanism unlikely. The solution of solar-nebula Ar in an enstatite melt²⁰ is also unlikely, because it would have required very high gas pressures during chondrule formation to compensate for the low solubility. □

Methods

Extraction efficiencies of the laser system were determined by comparing noble-gas concentrations obtained by the laser-ablation method with those obtained by a conventional total-melting method. Laser pits (~200 shots) were made at even intervals over the whole surface of thin slices (S6 and S8) as a check pattern, and gave a modal abundance of ablated minerals similar to that of each whole rock. The same samples as used for the laser-ablation method were heated at 1,800 °C for 20 min in a Mo crucible. The total-melting method guarantees complete degassing and thus no elemental fractionation. Based on the experiment, we estimated extraction efficiencies of 1.39, 1.09 and 1.08 for ^{36}Ar , ^{84}Kr and ^{132}Xe , respectively. Though a slight increase in the extraction efficiency for the lighter noble gas ^{36}Ar was observed, the mass releasing ^{36}Ar is only 40% larger than the melted material in each laser pit. Hence the spatial resolution of the laser system for ^{36}Ar is only 1.2 times as large as the apparent diameter of the laser pit—typically 50–80 μm .

Received 9 March; accepted 9 July 2001.

- Rubin, A. E. Petrologic, geochemical and experimental constraints on models of chondrule formation. *Earth Sci. Rev.* **50**, 3–27 (2000).
- Nakamura, T., Nagao, K. & Takaoka, N. Microdistribution of primordial noble gases in CM chondrites determined by in situ laser microprobe analysis: Decipherment of nebular processes. *Geochim. Cosmochim. Acta* **63**, 241–255 (1999).
- Kim, J. S. & Marti, K. Distribution of some highly volatile elements in chondrules. *Meteoritics* **29**, 482 (1994).
- Miura, Y. N. & Nagao, K. *Antarctic Meteorites* Vol. 22, 118–120 (National Inst. Polar Res., Tokyo, 1997).
- Crabb, J. & Anders, E. Noble gases in E-chondrites. *Geochim. Cosmochim. Acta* **45**, 2443–2464 (1981).
- Wacker, J. F. & Marti, K. Noble gas components in clasts and separates of the Abee meteorite. *Earth Planet. Sci. Lett.* **62**, 147–158 (1983).
- Eberhardt, P. et al. Trapped solar wind noble gases in Apollo 12 lunar fines 12001 and Apollo 11 breccia 10046. *Proc. Lunar Planet. Sci. Conf.* **III**, 1821–1856 (1972).
- Okazaki, R. *Origin of Noble Gases in Enstatite Chondrites and Ureilites: Evolutionary Processes of Meteoritic Materials in the Early Solar System*. Thesis, Kyushu Univ. (2000).
- Lewis, R. S., Srinivasan, B. & Anders, E. Host phases of a strange xenon component in Allende. *Science* **190**, 1251–1262 (1975).
- Wieler, R., Anders, E., Baur, H., Lewis, R. S. & Signer, P. Characterisation of Q-gases and other noble gas components in the Murchison meteorite. *Geochim. Cosmochim. Acta* **56**, 2907–2921 (1992).
- Crabb, J. & Anders, E. On the siting of noble gases in E-chondrites. *Geochim. Cosmochim. Acta* **46**, 2351–2361 (1982).
- Kimura, M. & Lin, Y. Petrological and mineralogical study of enstatite chondrites with reference to their thermal histories. *Antarct. Meteorit. Res.* **12**, 1–18 (1999).
- Okazaki, R., Takaoka, N., Nakamura, T. & Nagao, K. Cosmic-ray exposure ages of enstatite chondrites. *Antarct. Meteorit. Res.* **13**, 153–169 (2000).
- Hewins, R. H. in *Chondrules and the Protoplanetary Disk* (eds Hewins, R. H., Jones, R. H. & Scott, E. R. D.) 257–264 (Cambridge Univ. Press, 1996).
- Shu, F. H., Shang, H., Gounelle, M., Glassgold, A. E. & Lee, T. The origin of chondrules and refractory inclusions in chondritic meteorites. *Astrophys. J.* **548**, 1029–1050 (2001).
- Hiyagon, H. *Preliminary Studies on Partition of Rare Gases between Crystals and Melts*. Thesis, Univ. Tokyo (1981).
- Goswami, J. N., Lal, D. & Wilkening, L. L. Gas-rich meteorites: Probes for particle environment and dynamical processes in the inner solar system. *Space Sci. Rev.* **37**, 111–159 (1984).
- Edwards, S., Ray, T. & Mundt, R. in *Protostars and Planets III* (eds Levy, E. H. & Lunine, J. I.) 567–602 (Univ. Arizona Press, Tucson, 1993).
- Dodd, R. T. *Meteorites: a Petrologic-chemical Synthesis* (Cambridge Univ. Press, 1981).
- Kirsten, T. Incorporation of rare gases in solidifying enstatite melts. *J. Geophys. Res.* **73**, 2807–2810 (1968).
- Anders, E. & Grevesse, N. Abundances of the elements: Meteoritic and solar. *Geochim. Cosmochim. Acta* **53**, 197–214 (1989).
- Ozima, M. & Podosek, F. A. *Noble Gas Geochemistry* (Cambridge Univ. Press, 1983).
- Pepin, R. O. On the origin and early evolution of terrestrial planet atmospheres and meteoritic volatiles. *Icarus* **92**, 2–79 (1991).

Correspondence and requests for materials should be addressed to R.O.
(e-mail: okazaki@eqchem.s.u-tokyo.ac.jp).

Fast heating of ultrahigh-density plasma as a step towards laser fusion ignition

R. Kodama*, P. A. Norreys†, K. Mima*, A. E. Dangor‡, R. G. Evans§, H. Fujita*, Y. Kitagawa*, K. Krushelnick‡, T. Miyakoshi*, N. Miyanaga*, T. Norimatsu*, S. J. Rose†, T. Shozaki*, K. Shigemori*, A. Sunahara*, M. Tampon*, K. A. Tanaka*||, Y. Toyama*, T. Yamanaka* & M. Zepf‡

* Institute of Laser Engineering, Osaka University, 2-6 Yamada-oka, Suita Osaka 565-0871, Japan

† Rutherford Appleton Laboratory, Chilton, Didcot OX11 0QX, UK

‡ Blackett Laboratory, Imperial College, London SW7 2BZ, UK

§ Department of Physics, University of York, Heslington, York YO1 5DD, UK

|| Faculty of Engineering, Osaka University, 2-6 Yamada-oka, Suita Osaka 565-0871, Japan

Modern high-power lasers can generate extreme states of matter that are relevant to astrophysics¹, equation-of-state studies² and fusion energy research^{3,4}. Laser-driven implosions of spherical polymer shells have, for example, achieved an increase in density of 1,000 times relative to the solid state⁵. These densities are large

enough to enable controlled fusion, but to achieve energy gain a small volume of compressed fuel (known as the 'spark') must be heated to temperatures of about 10^8 K (corresponding to thermal energies in excess of 10 keV). In the conventional approach to controlled fusion, the spark is both produced and heated by accurately timed shock waves⁴, but this process requires both precise implosion symmetry and a very large drive energy. In principle, these requirements can be significantly relaxed by performing the compression and fast heating separately^{6–10}; however, this 'fast ignitor' approach⁷ also suffers drawbacks, such as propagation losses and deflection of the ultra-intense laser pulse by the plasma surrounding the compressed fuel. Here we employ a new compression geometry that eliminates these problems; we combine production of compressed matter in a laser-driven implosion with picosecond-fast heating by a laser pulse timed to coincide with the peak compression. Our approach therefore permits efficient compression and heating to be carried out simultaneously, providing a route to efficient fusion energy production.

In order to heat the compressed matter separately, the heating energy needs to be deposited on a timescale of less than 10^{-11} s, as the compression disassembles on this timescale. This is the fast ignitor⁷ approach. Present-day short-pulse laser technology is, in principle, capable of delivering sufficient energy in the required timescale, the largest laser to date having a peak power of 10^{15} W and pulse durations of 10^{-12} s (ref. 11). The laser energy is coupled to the highly compressed plasma via relativistic electrons that are efficiently generated when such an ultra-intense laser interacts with a high-density plasma^{11–21}. The extremely large electromagnetic fields of the laser accelerate the electrons into the high-density matter. (For a laser with a wavelength of 1 μ m, the typical conversion efficiency to the relativistic electrons has been measured to be about 30–40% at intensities above 10^{19} W cm⁻² (refs 10, 21).) The electrons then propagate to the high-density region where they deposit their energy.

We have measured the propagation of such relativistic electrons in a solid, and the associated heating effects, by examining the interactions of a 40-TW/0.5-ps laser pulse with a solid aluminium target. Images (obtained using ultraviolet light; Fig. 1) showing heated regions were temporally separated from any other hydrodynamic heating such as shock-wave and/or heat-wave propagation

processes by the use of a two-dimensional (2D) spatially resolved high-speed sampling camera²². The heating images indicate the propagation of the high-density electrons and collimation with a divergence (full-width at half-maximum) of 20–30°. The propagation of the large relativistic electron current is made possible by a return current of colder electrons that compensates the relativistic current almost perfectly. Magnetic fields associated with the current flow also serve to keep the electrons flowing initially in a narrow filament of the order of the laser spot diameter^{23–25}.

The experiments on heating of ultrahigh-density plasmas were performed on the Gekko XIII laser at the University of Osaka. This laser has 12 beams for nanosecond pulses, with a maximum energy of 15 kJ at 0.53 μ m wavelength, and a synchronized subpicosecond-pulse beam with a power of 100 TW and a pulse energy of 60 J (ref. 26). Conventional laser fusion experiments are conducted with spherically symmetrical targets to achieve high densities and the formation of the spark. This geometry was also envisaged in the original fast ignitor proposal⁷. Potential problems with this approach are propagation losses and deflection²⁷ of the ultra-intense laser pulse in the plasma surrounding the highly compressed plasma, and the transport of the relativistic electron beam through a substantial length of a plasma²⁸. Here we describe an experiment that departs from the original arrangement by inserting a gold cone (with an opening angle of 60°) into the shell (Fig. 2a; and S. Hatchett, unpublished work). The cone is designed to keep the propagation path of the short-pulse laser free from the plasma that forms around the imploding shell, thereby completely avoiding laser propagation issues. The cone tip was set at 50 μ m from the centre of the shell, and ensures that the compressed dense plasma forms at the tip of the cone while leaving the cone intact. The proximity between the cone tip and the core plasma also reduces the sensitivity to electron-beam propagation instabilities and losses. The cone walls are on a radius from the centre of the shell, minimizing disruptions to the spherical symmetry of the implosion so that high densities can be achieved. The laser energy used for compression here was restricted to 1.2 kJ in 1-ns-long pulses to ensure that the internal energy of the core after compression (which is about 5% efficient) was of similar magnitude to the short-pulse laser energy available (~ 60 J). This facilitates the measurement of the core plasma heating.

Figure 2b is a typical X-ray image of the implosion of the

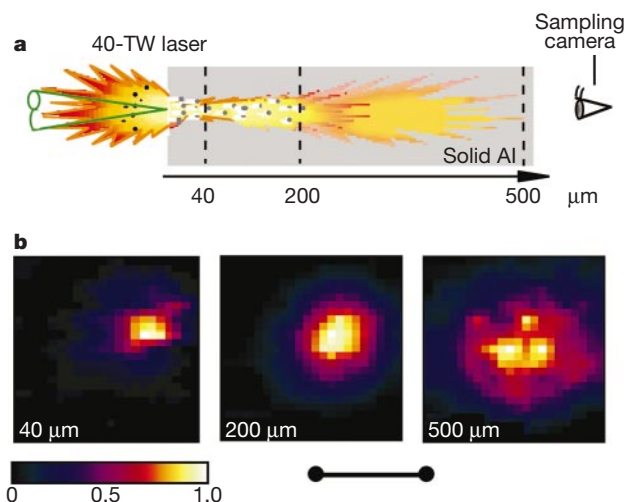


Figure 1 Ultraviolet images showing the heating of solid targets by relativistic electrons, and a sketch of the set-up used to obtain the images. **a**, 40-TW/20-J laser light was interacted with the Al solid target to create the electrons. The electrons heated the rear of the target, and this side of the target was imaged with the high-speed sampling camera. **b**, The heating images were obtained with the sampling camera for targets of different

thickness (from 10 μ m to 500 μ m). The scale bar below the images corresponds to 200 μ m on the target. The colour bar presents a linear intensity scale of the emission normalized by the peak intensity for each of the targets, to show clearly the difference in pattern. The relative peak intensities are about 5, 1 and 0.8 for the 40- μ m, 200- μ m and 500- μ m targets, respectively.

deuterated-polystyrene (CD) shell without injection of the short-pulse laser. The imploded core plasma was created at the centre of the unimploded shell, close to the tip of the cone. The compressed density was estimated by time-resolved radiographic measurements—the compressed plasma was illuminated with X-rays from a secondary target, and the images were recorded on a multi-frame high-speed camera (100 ps per frame). The core plasma size was measured in this way to be 40–45 μm in diameter. In addition, the area corresponding to a density of 0.1–0.5 g cm^{-3} was obtained from the reduced brightness of the backlighting radiation surrounding the core using the calculated opacity (Fig. 2c). We know the initial mass of the shell target is 4×10^{-6} g, the mass of the absorbing area is 2.8×10^{-7} g from the opacity measurement, and the remaining

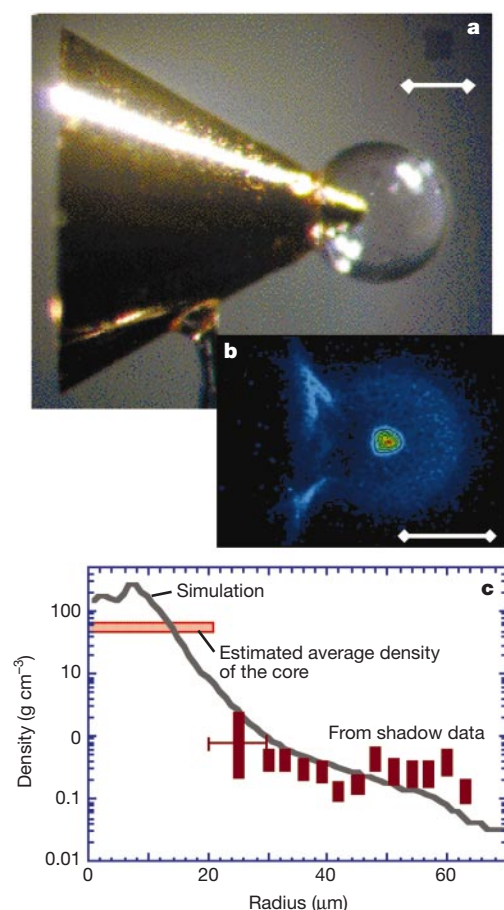


Figure 2 The implosion target for efficient heating of the highly compressed plasma, an X-ray image of the implosion, and the density profile of the plasma. The scale bars correspond to 250 μm on the target. **a**, A gold cone is attached to a deuterated-polystyrene (CD) shell (500 μm diameter, 7 μm wall thickness). 9 laser beams with a 1-ns duration are used to implode the shell at the tip of the cone. The 100-TW short-pulse laser is injected from the open side of the cone. **b**, Typical X-ray image showing the well imploded core plasma on the tip of the cone. The dimmer features correspond to the outline of the cone (left). The halo concentric with the bright core feature is emitted during the acceleration phase of the shell. **c**, Density profiles of the compressed plasma from the X-ray shadow, and a two-dimensional (2D) hydrodynamic simulation. The densities from the shadow are obtained by taking into account the opacity calculation. The errors shown are due to the spatial resolution. The 2D simulation code²⁹ is coupled with a one-dimensional (1D) simulation code including all the important physics. The initial conditions of the density and temperature given at the laser plasma interaction phase from the 1D code are introduced to the 2D code. In the code, incomplete spherical shock convergence and the interaction of the shell with the cone are treated, including long-wavelength hydrodynamic instability processes. However, no account is taken of shorter-wavelength perturbations of the shell caused at the acceleration phase.

mass after long-pulse laser ablation is 2.2×10^{-6} g (the ablated mass is 1.5×10^{-6} g) from both simulations and the experimental database on the mass ablation rate. From these, we estimate the average core density as 50–70 g cm^{-3} for the 40–45 μm core diameter. Given the non-uniformity of the laser illumination, the inferred density is consistent with that calculated using two-dimensional hydrodynamic simulations (an average density of 80–100 g cm^{-3} ; Fig. 2c). The simulation also suggests that the insertion of the cone only marginally reduces the achieved compressed density (by about 20–30%) compared with full spherical implosion. This shows that implosions with a cone insert are compatible with achieving the high densities required for achieving fusion gain, while at the same time allowing for reliable and efficient coupling of the laser to the highly compressed plasma.

We investigated the heating efficiency by injecting the 100-TW heating laser into the cone at the moment of maximum compression. The electron beam created by the 100-TW laser was measured with electron spectrometers during interactions with the cone target; we found that kT for the beam was 2–3 MeV. The conversion efficiency was obtained using the same laser with a plane target at a similar intensity ($10^{19} \text{ W cm}^{-2}$); the energy of the electron beam was 18–24 J, representing 30–40% of the laser energy. To check the shape and divergence of the heated region associated with the electron beam in this new geometry, the 100-TW laser was injected into an identical cone target, but the tip of the cone touched an Al block (with a thickness of 200 μm at the point of contact), rather than being embedded in the CD shell. The heating area was elliptical in shape ($40 \times 25 \mu\text{m}^2$; Fig. 3a), and the beam divergence was less than 20° from the size ($130 \times 80 \mu\text{m}^2$) of the heated region on the back of the 200- μm Al. Figure 3b shows an image sequence of the core emission with the heating pulse injected at the time of maximum compression in the cone insert geometry. Figure 3c shows a similar sequence, but this time the heating pulse arrives 150 ps after the peak compression. The heating due to the short-pulse beam is clearly visible in the sequence in Fig. 3c: the first peak of the core emission coincides with the peak compression, and then decreases before peaking again after the injection of the heating pulse. The size of the emission region due to the peak in compression is about 30 μm , and increases to 50 μm after the injection of the heating pulse—in good agreement with the independent measurement of the heating region using the Al block.

Heating of the highly compressed plasma was quantified by

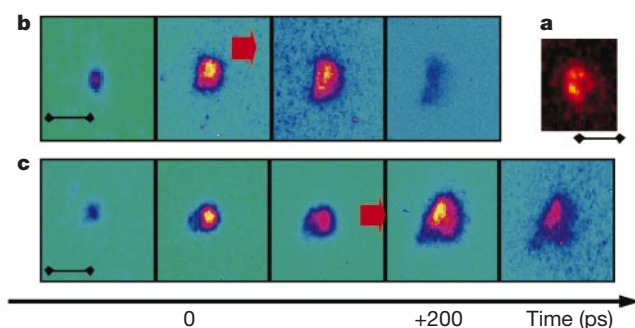


Figure 3 Time-integrated X-ray image of the short-pulse laser heating, and time-resolved X-ray images of the highly compressed plasma heated by the short-pulse laser. Time separation between each of the time-resolved images is 100 ps. The red arrows in the figure correspond to the timing of the short-pulse laser injection. The scale bar of the images shows 50 μm on the target. **a**, The image is observed from the beam-injection side of the cone attached to an Al block instead of the shell. **b**, The short-pulse laser was injected at a time close to the maximum compression of the shell. **c**, The injection timing is about 150 ps after the maximum compression. The heating by the short-pulse laser is temporally separated from the maximum compression heating of the shell.

Table 1 Neutron yield from the dense plasmas at different conditions

Target	E_i , no. of beams	E_s	N_f
Cone+shell	1.2 kJ, 9 beams	60 J	$(1-3) \times 10^5$
Cone+shell	1.2 kJ, 9 beams	0	$(0.8-1) \times 10^4$
Spherical shell only	2.6 kJ, 12 beams	0	$(2-3) \times 10^5$

Thermal neutron yields N_f from the highly compressed plasmas are listed for different targets, implosion laser energy E_i and heating energy from the short-pulse laser E_s . The E_i is nominal energy measured after the focusing lens and through a random phase plate. These neutron yields indicates temperatures of several hundred eV.

measuring the increase in production of thermonuclear neutrons. Neutrons are generated by the fusion of two deuterium nuclei to a ^3He nucleus ($\text{d}(\text{d},\text{n})^3\text{He}$) in the imploded plasma, and provide a precise measurement of the plasma temperature. The neutron energy spectra were obtained using time-of-flight scintillator/photomultiplier detectors from two different angles. Peaks at an energy of 2.45 MeV are observed, corresponding to neutrons from a thermonuclear $\text{d}(\text{d},\text{n})^3\text{He}$ reaction. The neutron time-of-flight spectrum in Fig. 4 shows a signal corresponding to a thermonuclear neutron yield of $(2 \pm 1) \times 10^5$ neutrons, and was taken when the heating pulse was injected at maximum compression. This neutron yield was more than 10 times the numbers $((9 \pm 1) \times 10^3)$ observed when no heating pulse was present or when the heating pulse was not timed to coincide with maximum compression. In order to replicate the neutron yield achieved at optimal timing in the conventional fashion (no heating pulse and a spherically symmetric implosion), a laser energy of 2.6 kJ was required to drive the implosion. Therefore the total energy required to achieve the observed neutron yield has been reduced by half, which is a clear demonstration of the increased efficiency that can be achieved by separating the compression and the heating phase in laser fusion experiments. These results, summarized in Table 1, provide (to our knowledge) the first clear evidence of effective heating of compressed plasma using an ultrahigh-intensity, short-pulse laser.

The efficiency of the energy coupling of the energetic electrons to the highly compressed plasma can be estimated from the neutron yield and the heated volume inferred from the X-ray images. In order to increase the neutron yield by a factor of 10–30, a temperature increase of about 120 eV ($\pm 20\%$) is required for initial temperature regions of several hundred eV. The heated mass is estimated from the density ($50-70 \text{ g cm}^{-3}$), and the volume is estimated from the size of the electron beam ($25 \times 40 \mu\text{m}^2$) and the length of the high-density plasma ($40 \mu\text{m}$). Taking account of the beam divergence, the volume might be 1.2–1.3 times larger than this assumption. The heated distance might also be $30 \mu\text{m}$ from the

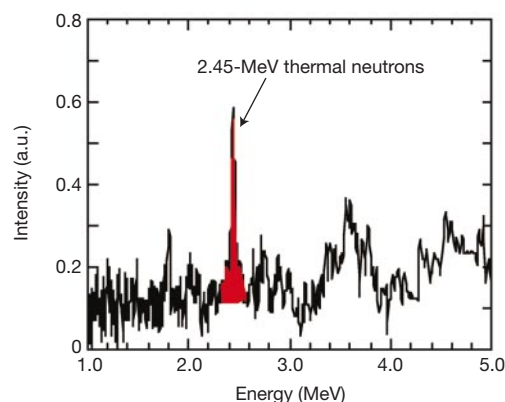


Figure 4 Neutron spectrum from the highly dense plasma heated by the short-pulse laser at the time of maximum compression. The peak at 2.45 MeV corresponds to neutrons from the thermonuclear deuterium–deuterium fusion reaction. The signal to noise ratio is ≥ 5 .

core emission instead of $40 \mu\text{m}$. Then the volume assumption will include an error of about $\pm 30\%$. To heat uniformly the volume of density $50-70 \text{ g cm}^{-3}$ ($25 \times 40 \times 40 \mu\text{m}^3$) requires 12–16 J of the energetic electrons produced by the short-pulse laser. The total coupling efficiency of this laser to the compressed dense plasma could therefore be 20–27%, having estimation errors of $\pm 8\%$ (12–35%) from the temperature and from the heated volume. This is an encouraging result.

Using the minimum 20% coupling efficiency observed in this experiment, we estimate the short-pulse laser energy needed to create a sufficiently large spark to ignite a deuterium–tritium (DT) fusion pellet ($T = 12 \text{ keV}$, $\rho = 600 \text{ g cm}^{-3}$ and $\rho r = 0.4-0.6 \text{ g cm}^{-2}$) to be 10–20 kJ (ref. 6), which seems feasible. Of course, we are still uncertain how the increase in the electron beam current will affect the propagation and energy deposition in the highly compressed DT plasma for a full-scale fusion experiment. For example, a fourfold increase in energy concentration is required as compared with this first demonstration experiment (as a smaller, $16\text{-}\mu\text{m}$ -diameter core needs to be heated due to the higher compression needed for ignition). Another issue to be resolved for the future ignition experiments is the fabrication of a uniform cryogenic fuel layer, such as a foam shell filled with liquid DT fuel. Nevertheless, we emphasized that the temperature of the energetic electrons used in our experiment is closely matched to the requirements of a full-scale fusion experiment. □

Received 4 May; accepted 6 July 2001.

- Remington, B. A., Arnet, D., Drake, R. P. & Takabe, H. Modeling astrophysical phenomena in the laboratory with intense lasers. *Science* **284**, 1488–1493 (1999).
- Ichimaru, S. & Kitamura, H. Pycnonuclear reactions in dense astrophysical and fusion plasmas. *Phys. Plasmas* **6**, 2649–2671 (1999).
- Nuckolls, J., Wood, L., Thiessen, A. & Zimmerman, G. Laser compression of matter to super-high densities. *Nature* **239**, 139–142 (1972).
- Lindl, J., McCrory, R. L. & Campbell, E. M. Progress toward ignition and burn propagation in inertial confinement fusion. *Phys. Today* **45**, 32–40 (1992).
- Azechi, H. et al. High density compression experiments at ILE, Osaka. *Laser Part. Beams* **9**, 193–207 (1991).
- Piriz, A. R. & Sanchez, M. M. Analytical model for the dynamics of fast ignition. *Phys. Plasmas* **5**, 2721–2726 (1998).
- Tabak, M. et al. Ignition and high gain with ultrapowerful lasers. *Phys. Plasmas* **1**, 1626–1634 (1994).
- Atzeni, S. Inertial fusion fast ignitor: Igniting pulse parameter window vs the penetration depth of the heating particles and the density of the precompressed fuel. *Phys. Plasmas* **6**, 3316–3326 (1999).
- Norreys, P. et al. Experimental studies of the advanced fast ignitor scheme. *Phys. Plasmas* **7**, 3721–3726 (2000).
- Kodama, R. et al. Fast ignition research at the institute of laser engineering Osaka University. *Phys. Plasmas* (in the press).
- Perry, M. D. & Mourou, G. Terawatt to petawatt subpicosecond lasers. *Science* **264**, 917–924 (1994).
- Kruer, W. E. and Estabrook, K. JxB heating by very intense laser light. *Phys. Fluids* **28**, 430–432 (1985).
- Bruneel, F. Not-so-resonant, resonant absorption. *Phys. Rev. Lett.* **59**, 52–55 (1987).
- Lefebvre, E. & Bonnaud, G. Transparency/opacity of a solid target illuminated by an ultrahigh-intensity laser pulse. *Phys. Rev. Lett.* **74**, 2002–2005 (1995).
- Malka, G. & Miquel, J. L. Experimental validation of the linear theory of stimulated Raman scattering driven by a 500-fs laser pulse in a preformed underdense plasma. *Phys. Rev. Lett.* **74**, 4655–4658 (1996).
- Pukhov, A. & Meyer-ter-Vehn, J. Laser hole boring into overdense plasma and relativistic electron currents for fast ignition of ICF targets. *Phys. Rev. Lett.* **79**, 2686–2689 (1997).
- Key, M. H. et al. Hot electron production and heating by hot electrons in fast ignitor research. *Phys. Plasmas* **5**, 1966–1972 (1998).
- Kodama, R. et al. Long-scale jet formation with specularly reflected light in ultraintense laser-plasma interactions. *Phys. Rev. Lett.* **84**, 674–677 (2000).
- Santala, M. I. K. et al. Effect of the plasma density scale length on the direction of fast electrons in relativistic laser-solid interactions. *Phys. Rev. Lett.* **84**, 1459–1463 (2000).
- Tanaka, K. A. et al. Studies of ultra-intense laser plasma interactions for fast ignition. *Phys. Plasmas* **7**, 2014–2022 (2000).
- Wharton, K. B. et al. Experimental measurements of hot electrons generated by ultraintense ($>10^{19} \text{ W cm}^{-2}$) laser plasma interactions on solid-density targets. *Phys. Rev. Lett.* **81**, 822–825 (1998).
- Kodama, R. et al. Development of a two-dimensional space-resolved high speed sampling camera. *Rev. Sci. Instrum.* **70**, 625–628 (1999).
- Davies, J. R., Bell, A. R. & Tatarakis, M. Magnetic focusing and trapping of high-intensity laser generated fast electrons at the rear of solid targets. *Phys. Rev. E* **59**, 6032–6036 (1999).
- Tatarakis, M. et al. Plasma formation on the front and rear of plastic targets due to high-intensity laser-generated fast electrons. *Phys. Rev. Lett.* **81**, 999–1002 (1998).
- Honda, M., Meyer-ter-Vehn, J. & Pukhov, A. Collective stopping and ion heating in relativistic-electron-beam transport for fast ignition. *Phys. Rev. Lett.* **85**, 2128–2131 (2000).
- Kato, Y. et al. Fast ignition and related plasma physics issues with high-intensity lasers. *Plasma Phys. Control. Fusion* **39**, 145–151 (1997).

27. Duda, B. J., Hemker, R. G., Tzeng, K. C. & Mori, W. B. A long-wavelength hosing instability in laser-plasma interactions. *Phys. Rev. Lett.* **83**, 1978–1981 (1999).
28. Hain, S., Cornolti, F. & OPOWER, H. Hydrodynamic models and schemes for fast ignition. *Laser Part. Beams* **17**, 245–263 (1999).
29. Sunahara, A., Takabe, H. & Mima, K. 2D simulation of hydrodynamic instability in ICF stagnation phase. *Fusion Eng. Design* **44**, 163–169 (1999).

Acknowledgements

We thank the mm-Wave Technology Centre at the Rutherford Appleton Laboratory, and the target fabrication, laser operation and data acquisition groups at ILE Osaka University. This work was supported by the Japan Society for the Promotion of Science, and the UK Royal Society.

Correspondence and requests for materials should be addressed to R.K. (e-mail: ryo@ile.osaka-u.ac.jp).

Formation of ordered ice nanotubes inside carbon nanotubes

Kenichiro Koga*, G. T. Gao†‡, Hideki Tanaka§ & X. C. Zeng†

* Department of Chemistry, Fukuoka University of Education, Fukuoka 811-4192, Japan

† Department of Chemistry and Center for Materials and Analysis, University of Nebraska, Lincoln, Nebraska 68588, USA

§ Department of Chemistry, Okayama University, 3-1-1, Tsushima, Okayama 700-8530 Japan

Following their discovery¹, carbon nanotubes have attracted interest not only for their unusual electrical and mechanical properties, but also because their hollow interior can serve as a nanometre-sized capillary^{2–7}, mould^{8–11} or template^{12–14} in material fabrication. The ability to encapsulate a material in a nanotube also offers new possibilities for investigating dimensionally confined phase transitions¹⁵. Particularly intriguing is the conjecture¹⁶ that matter within the narrow confines of a carbon nanotube might exhibit a solid–liquid critical point¹⁷ beyond which the distinction between solid and liquid phases disappears. This unusual feature, which cannot occur in bulk material, would allow for the direct and continuous transformation of liquid matter into a solid. Here we report simulations of the behaviour of water encapsulated in carbon nanotubes that suggest the existence of a variety of new ice phases not seen in bulk ice, and of a solid–liquid critical point. Using carbon nanotubes with diameters ranging from 1.1 nm to 1.4 nm and applied axial pressures of 50 MPa to 500 MPa, we find that water can exhibit a first-order freezing transition to hexagonal and heptagonal ice nanotubes, and a continuous phase transformation into solid-like square or pentagonal ice nanotubes.

Carbon nanotubes can be wetted by liquids⁴ whose surface tension does not exceed about 200 mN m^{−1}. Thus, in principle, pure water can be drawn into open-ended nanotubes by capillary suction⁵. Once inside, water molecules are expected to form quasi-one-dimensional (Q1D) structures that might form new phases of ice, different from the 13 polymorphic phases of bulk ice identified experimentally thus far¹⁸. We carried out molecular dynamics (MD) simulations at constant temperature (T) and axial-pressure (P_{zz}) of water confined within ‘armchair’¹⁹ (R,R) single-walled carbon nanotubes (SWCNs). We used nanotubes with indices $R = 14–18$, corresponding to tubes with diameters of 11.1, 11.9, 12.6, 13.4 and 14.2 Å, respectively. The phase behaviour of the confined water was

examined in several series of the MD simulations, each series corresponding to an isobaric path or an isothermal path in the P_{zz} – T phase diagram at a given R (see Methods for details).

The first series of simulations follows an isobaric path of 50 MPa. The temperature was lowered stepwise starting from 320 K or higher, where the system is in a liquid state, to 240 K or lower. The potential energy of water in each type of SWCN is plotted in Fig. 1. In the wide SWCNs (16,16) and (17,17), the potential energy drops abruptly (Fig. 1c and d) on cooling and jumps sharply on heating. This marked hysteresis-loop behaviour signifies a first-order phase transition. Structural analysis reveals that the low- T phase is a Q1D n -gonal ‘ice nanotube’ composed of orderly stacked n -membered water rings²⁰, where $n = 6$ (hexagonal) in (16,16) and $n = 7$ (heptagonal) in (17,17) SWCNs. In both types of nanotube, the molar volume of water decreases during the liquid-to-ice nanotube transition; that is, the confined water shrinks on freezing. In the widest SWCN (18,18), however, crystallization was not observed within the timescale of simulation. In the narrower SWCNs (14,14) and (15,15), the potential energy also drops markedly on cooling below 300 K, but the change is not as sharp as in the wider nanotubes. Structural analysis shows that confined water has liquid-like disordered structure at high T but turns into solid-like ordered structure at low T —a square nanotube in (14,14) and a pentagonal nanotube in (15,15) SWCN. At 240 K, the calculated diffusion constants (along the axial direction) are $D = 3 \times 10^{-10}$ cm² s^{−1} in the (14,14) SWCN, and $D < 1 \times 10^{-10}$ cm² s^{−1} in the (15,15) SWCN, which are comparable to D of bulk ice²¹. At 300 K, $D = 1 \times 10^{-5}$ cm² s^{−1} and $D = 2 \times 10^{-5}$ cm² s^{−1} respectively. More interestingly, besides the less sharp change in the potential energy, the hysteresis loop was not observed in the cooling and heating process, a signature of continuous transformation from liquid-like to solid-like state of water.

In real-world experiments, the atomic structures of Q1D crystals can be determined by using transmission electron microscopy¹¹. Simulations provide this information directly. Figure 2 displays snapshots of the Q1D n -gonal ($n = 4–6$) ice nanotubes and the corresponding Q1D liquid phases inside the (14,14), (15,15) and

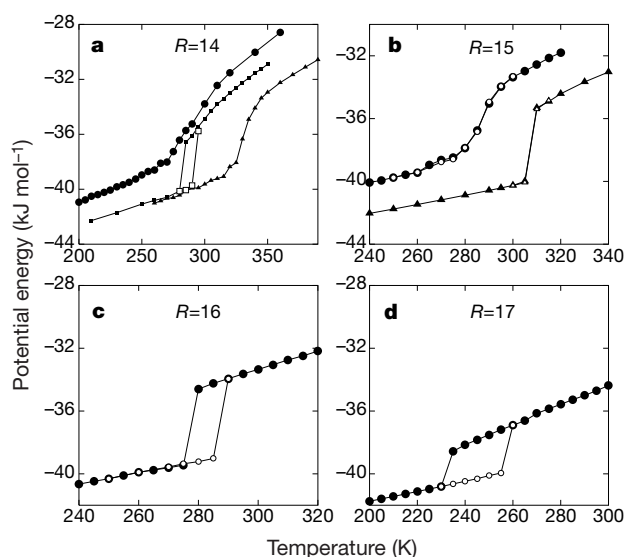


Figure 1 Potential energy against temperature for water confined in four types of single-walled carbon nanotube. The nanotubes are armchair (R,R) SWCNs, where $R = 14–17$ (a–d, respectively). The potential energy is due to the water–water intermolecular interactions, and the water–SWCN interaction energy is excluded. The applied axial pressure is 50 MPa (circles), 200 MPa (squares), and 500 MPa (triangles). Filled and unfilled symbols indicate the cooling and heating process, respectively.

‡ Present address: Department of Chemistry, U.S. Naval Academy, Annapolis, Maryland 21402, USA

(16,16) SWCNs, showing that the disordered liquid-like structures are transformed at low T into distinct n -gonal nanotube structures with long-range order in the axial direction. The n -gonal ice nanotubes satisfy the 'bulk ice rule': every water molecule serves as a double donor and a double acceptor of hydrogen bonds, and every water molecule is hydrogen-bonded to exactly four nearest-neighbour molecules. However, a few residual defects might occur in the ice nanotubes, and this would result in interesting helical ice nanotube structure (shown in Fig. 2b). An independent chemical-potential calculation²² indicates that bulk water at a pressure of 83 MPa is in equilibrium with a hexagonal ice at $P_{zz} = 50$ MPa in the (16,16) SWCN at 260 K. The difference between external and internal pressures suggests that water can be easily drawn into the SWCNs^{4,5} and, under moderate compression, form a hexagonal ice nanotube.

A comparison of the snapshot of the hexagonal ice nanotube with that of the liquid in the (16,16) SWCN (Fig. 2c and f) shows that the solid phase has a hollow-tube structure whereas the liquid phase has not. The radial density profile of water in the (16,16) SWCN accordingly has a distribution peak at $r = 0$ (axis of the SWCN) at higher T . This peak disappears abruptly between 275 and 280 K (Fig. 3a); that is, the system changes its 'symmetry' characteristics markedly on freezing. The abrupt symmetry change indicates that

freezing in wider carbon nanotubes is a first-order transition. This conclusion is supported by the curves of chemical potential against temperature for the confined liquid and hexagonal ice nanotube under 50 MPa, respectively (Fig. 3b). The two curves intersect at about 280 K, the melting point of the hexagonal ice nanotube, which falls within the region enclosed by the hysteresis loop shown in Fig. 1c. In the narrower (14,14) and (15,15) SWCNs, the liquid phase has a hollow-tube structure (Fig. 2d and e), in sharp contrast to the liquid phase in (16,16) SWCN. The 4-membered water-ring structure seen in the liquid phase in (14,14) SWCN resembles that of the square ice nanotube. This similarity in the structural characteristics (squareness and hollowness) of the two phases explains why the confined liquid water phase can transform continuously into the solid phase on cooling. Similar arguments apply to the pentagonal counterpart in (15,15) SWCN.

We also examined the phase behaviour at higher P_{zz} in the (14,14) SWCN. The potential energy on the 500 MPa isobar decreases smoothly with a distinct change between 320 and 340 K on cooling (Fig. 1a). On the 200 MPa isobar, however, the potential energy drops sharply at 280 K and a marked hysteresis appears on heating. On both high- P_{zz} isobars the low- T phase is the pentagonal nanotube. That is, liquid water undergoes a continuous transformation into the square-nanotube phase at 50 MPa and into the pentagonal-nanotube phase at 500 MPa, but undergoes a first-order transition to the pentagonal nanotube at 200 MPa. Figure 4a and b shows the axial and angular order parameters for square ($n = 4$) and pentagonal ($n = 5$) phases, defined by $\langle |\Sigma_j \exp(2\pi i N z_j / n l)|^2 / N^2 \rangle$ and $\langle |\Sigma_j \exp(i n \theta_j)|^2 / N^2 \rangle$, where z_j and θ_j are the z -coordinate and the angle of j -th molecules in the cylindrical coordinate system, l is the length of the carbon nanotube, and $\langle \dots \rangle$ denotes the average. As expected, the two order parameters for the pentagonal phase show a sharp increase as T is lowered only when $P_{zz} = 200$ MPa.

The phase behaviour of water in the (14,14) SWCN deserves special attention because it implies the possible existence of a solid-liquid critical point, which cannot occur in bulk systems¹⁷. To explore this possibility, we isothermally compressed the water stepwise from 50 to 900 MPa. In particular, long-time (200 ns) simulations were done for several P_{zz} from 250 to 330 MPa where all isotherms show a marked decline in potential energy and volume. The resulting potential-energy- P_{zz} and volume- P_{zz} isotherms are shown in Fig. 4c and d. On the 280 K isotherm, the potential energy and the volume show a sudden drop at 200 MPa on compression, and a sudden increase at 170 MPa on decompression; that is, a hysteresis loop appears. Structural analysis shows that the low-density phase is a square phase with a liquid-like disordered structure, whereas the high-density phase is a solid-like pentagonal-nanotube phase. The sudden changes in potential energy and volume evince a first-order transition between the two distinct

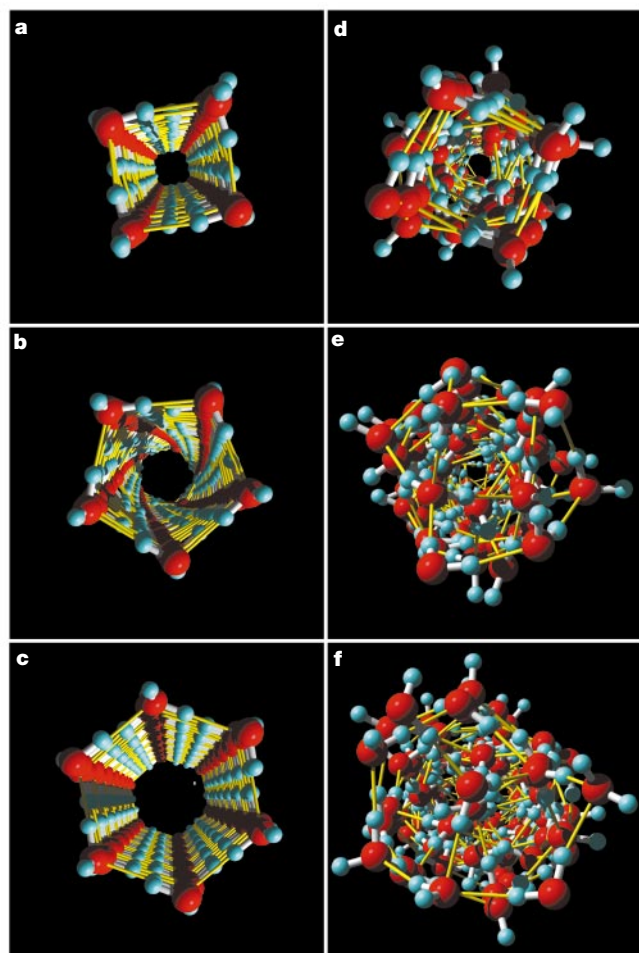


Figure 2 Snapshots of quenched molecular coordinates. **a**, Square; **b**, pentagonal; **c**, hexagonal ice nanotubes in (14,14), (15,15) and (16,16) SWCNs; **d** to **f**, the corresponding liquid phases. The ice nanotubes were formed on cooling under an axial pressure of 50 MPa in molecular dynamics simulations. The nearest-neighbour distances in both ice nanotube and encapsulated liquid water are fairly constant, about 2.7 to 2.8 Å, and this is in part responsible for the novel phase behaviour.

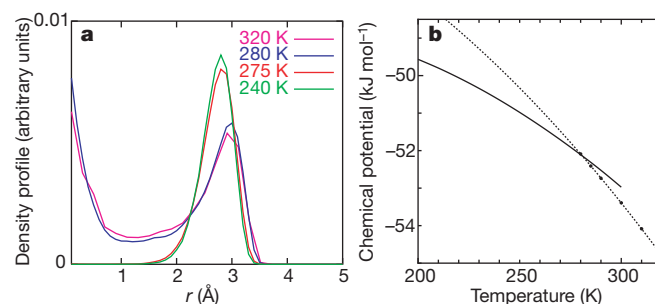


Figure 3 Properties associated with the first-order phase transition in the (16,16) SWCN at a fixed pressure of 50 MPa. **a**, Radial density profile of confined water at various temperatures. **b**, Chemical potential of liquid water (filled circles and dashed line) and the hexagonal ice nanotube (solid line) against temperature.

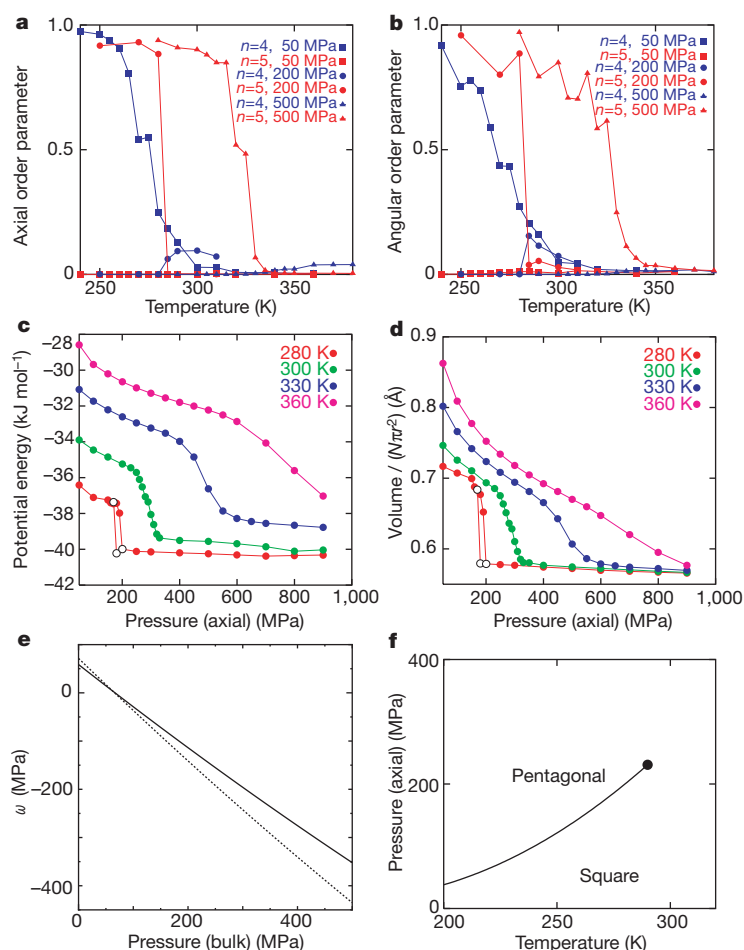


Figure 4 Properties associated with the phase transformation in the (14,14) SWCN. **a**, Axial; **b**, angular order parameters plotted against temperature for the square phase ($n = 4$) and pentagonal phase ($n = 5$). **c**, Curve of potential energy against pressure; **d**, volume against pressure. Filled and open marks indicate the compression and decompression process, respectively. **e**, Grand-potential density ω of the square ice nanotube (solid line) and pentagonal ice nanotube (dashed line) plotted against the

phases. On the compressing path, the diffusion constant drops abruptly from $0.9 \times 10^{-6} \text{ cm}^2 \text{ s}^{-1}$ at 190 MPa to $1.0 \times 10^{-10} \text{ cm}^2 \text{ s}^{-1}$ at 200 MPa, which reinforces the idea that the pentagonal-nanotube phase is a solid whereas the square phase is a liquid. On the 300 K isotherm, however, the transformation seems to be continuous as no abrupt jumps in potential energy and volume are found. On the 330 K isotherm, the changes in potential energy and volume become much less abrupt, and finally on the 350 K isotherm the changes are essentially smooth.

The first-order transitions observed in the MD simulation on the 200 MPa isobar and 280 K isotherm in the (14,14) SWCN indicate that the square liquid and pentagonal ice-nanotube phases must be separated by a first-order phase boundary. The grand-potential densities (see Methods) of the square- and pentagonal-ice nanotube at 260 K have an intersection at about 65 MPa, as plotted in Fig. 4e against the pressure of bulk water. This free-energy result, independent of the MD simulation, supports the existence of a first-order phase boundary separating the square- and pentagonal-ice nanotube phase. Figure 4f is a schematic T - P_{zz} phase diagram derived from the results of MD simulations and free-energy calculations. The solid line denotes the boundary of the first-order transition. It is known that on a temperature–pressure phase diagram a first-order phase boundary can go to infinity, intersect other phase boundaries, or end at a critical point. Here, it is most likely that the phase-

pressure P of bulk liquid water at 260 K. **f**, Schematic temperature/axial pressure phase diagram of water. The first-order phase boundary (solid line) that divides the pentagonal-nanotube phase from the square-nanotube phase in the low- T range gradually turns into a boundary that separates the pentagonal-nanotube phase from the square-liquid phase at high T . The first-order phase boundary is terminated by a critical point (the filled circle).

boundary line is terminated by a solid–liquid critical point beyond which the pentagonal ice-nanotube and the square liquid lose their identities, because the square-to-pentagonal nanotube transformation is continuous in the high- T and high- P_{zz} region. It appears from the isotherms in Fig. 4c and d that this critical point lies between 280 and 300 K, and 200 and 300 MPa. □

Methods

Molecular dynamics simulations

The TIP4P model²³ for water was used in the MD simulations, whose intermolecular potential energy is given by the sum of the long-range Coulomb potential and the short-range Lennard–Jones potential between the interaction sites. The force field of the model SWCN was taken to be a Lennard–Jones potential integrated over the cylindrical area of the SWCN using the area density of the carbon atoms and the potential parameters for graphite²⁴. A periodic boundary condition was applied in the axial direction of a tubule of length l containing 180 water molecules (182 when $R = 17$). The temperature T and the internal axial-pressure P_{zz} (the pressure tensor parallel to the axis) were maintained using a modified Nosé–Andersen's method²⁵. The axial dimension l fluctuates to keep P_{zz} constant during a simulation; continuous change in l is possible because the force field of the model SWCN is smooth and structureless, and thus is independent of l . The simulation time at each state point (given T , P_{zz} , and R) was 10–40 ns, and at certain states was over 200 ns, which is one or two orders of magnitude longer than a typical simulation time for bulk water. Equilibration at each state point was assured by the absence of any steady drift in the properties of interest, and by reproducibility of these properties in heating and cooling (or compressing and decompressing) processes, except for the hystereses due to the first-order transitions.

Structural analysis

Instantaneous configurations generated in the MD simulations were mapped onto corresponding potential-energy local-minimum configurations via the constant-volume steepest-descent method. These configurations were used for structural analysis.

Grand-potential density

Phase equilibria of water confined in an open-ended carbon nanotube in contact with bulk water at fixed chemical potential μ and temperature T are determined from the minima of grand-potential density, $\omega = (F - \mu N)/V$, where F and V are the Helmholtz free energy and the volume of the interior. The stable phase has the lowest ω and two phases coexist if they have an identical minimum. The relation between P and μ of bulk water was obtained from the free-energy calculation for bulk water²².

Free-energy calculation

For ice inside a carbon nanotube of length l , the Helmholtz free energy, $F(T, l)$, at temperature T is calculated as a sum of four contributions: (1) the potential energy of the inherent structure, $U_q(l)$; (2) the harmonic vibrational free energy, $f_j(T, l) = k_B T \sum_{i=1}^{3(N-1)} \ln(h\nu_i/k_B T)$ where k_B and h are the Boltzmann and Planck constants, and ν_i is the frequency of the i th normal mode; (3) the anharmonic vibrational free energy, $f_a(T, l) = T \int_0^1 [U_q(l) + 3Nk_B T' - U(T', l)/T'^2] dT'$, where $U(T', l)$ is the mean potential energy of ice at temperature T' ; and (4) the entropy term, TS_c where S_c is the configurational entropy arising from the disordered proton arrangement, given by $S_c = (N/n)k_B \ln 2$ for an n -gonal ice nanotube. Under a given internal axial-pressure, P_{zz} , the thermodynamic potential $G_{ice}(T, P_{zz}, l) = F(T, l) + P_{zz}Al$ has a minimum with respect to the variation of l for an arbitrarily defined inner area A . Thus, we can obtain the Gibbs free energy and the equilibrium length (l) as functions of P_{zz} , which generally differs from the pressure P of the bulk water. For liquid water in a SWCN, the interaction potential can be divided into the water–water interaction Φ_1 and the interaction of water molecules with carbon Φ_2 . The free energy is given by

$$F(T, l) = F_0(T, l) + \int_0^1 d\lambda (4\lambda^3 \Phi_1)$$

where $\langle \dots \rangle$ stands for the ensemble average of the system interacting with the potential ($\lambda^3 \Phi_1 + \Phi_2$) and $F_0(T, l)$ is the free energy of the ideal gas interacting only with carbon.

Received 21 December 2000; accepted 21 June 2001.

- Iijima, S. Helical microtubules of graphitic carbon. *Nature* **354**, 56–58 (1991).
- Pederson, M. R. & Broughton, J. Q. Nanocapillarity in fullerene tubules. *Phys. Rev. Lett.* **69**, 2689–2692 (1992).
- Ajayan, P. M. & Iijima, S. Capillarity-induced filling of carbon nanotubes. *Nature* **361**, 333–334 (1993).
- Dujardin, E., Ebbesen, T. W., Hiura, H. & Tanigaki, K. Capillarity and wetting of carbon nanotubes. *Science* **265**, 1850–1852 (1994).
- Tsang, S. C., Chen, Y. K., Harris, P. J. F. & Green, M. L. H. A simple chemical method of opening and filling carbon nanotubes. *Nature* **372**, 159–162 (1994).
- Ugarte, D., Chatelain, A. & de Heer, W. A. Nanocapillarity and chemistry in carbon nanotubes. *Science* **274**, 1897–1899 (1996).
- Kiang, C. H., Choi, J. S., Tran, T. T. & Bacher, A. D. Molecular nanowires of 1 nm diameter from capillary filling of single-walled carbon nanotubes. *J. Phys. Chem. B* **103**, 7449–7451 (1999).
- Ruoff, R. S. *et al.* Single crystal metals encapsulated in carbon nanoparticles. *Science* **259**, 346–348 (1993).
- Guerret-Picourt, C., Le Bouar, Y., Loiseau, A. & Pascard, H. Relation between metal electronic structure and morphology of metal compounds inside carbon nanotubes. *Nature* **372**, 761–764 (1994).
- David, V. P. *et al.* Controlled-size nanocapsules. *Nature* **374**, 602 (1995).
- Meyer, R. R. *et al.* Discrete atom imaging of one-dimensional crystals formed within single-walled carbon nanotubes. *Science* **289**, 1324–1326 (2000).
- Ajayan, P. M., Stephan, O., Redlich, P. & Colliex, C. Carbon nanotubes as removable templates for metal oxide nanocomposites and nanostructures. *Nature* **375**, 564–567 (1995).
- Dai, H. *et al.* Synthesis and characterization of carbide nanorods. *Nature* **375**, 769–772 (1995).
- Han, W., Fan, S., Li, Q. & Hu, Y. Synthesis of gallium nitride nanorods through a carbon nanotube-confined reaction. *Science* **277**, 1287–1289 (1997).
- Fan, X. *et al.* Atomic arrangement of iodine atoms inside single-walled carbon nanotubes. *Phys. Rev. Lett.* **84**, 4621–4624 (2000).
- Ball, P. New horizons in inner space. *Nature* **361**, 297 (1993).
- Stanley, H. E. in *Introduction to Phase Transitions and Critical Phenomena* 2 (Oxford Univ. Press, New York, 1971).
- Lobban, C., Finney, J. L. & Kuhs, W. F. The structure and ordering of ices III and V. *J. Chem. Phys.* **112**, 7169–7180 (2000).
- Hamada, N., Sawada, S. & Oshiyama, A. New one-dimensional conductors—graphitic microtubules. *Phys. Rev. Lett.* **68**, 1579–1581 (1992).
- Koga, K., Parra, R. D., Tanaka, H. & Zeng, X. C. Ice nanotubes: What does the unit cell look like? *J. Chem. Phys.* **113**, 5037–5040 (2000).
- Goto, K., Hondoh, T. & Higashi, A. Determination of diffusion coefficients of self-interstitials in ice with a new method of observing climb of dislocations by X-ray topography. *Jpn J. Appl. Phys.* **25**, 351–357 (1986).
- Gao, G. T., Zeng, X. C. & Tanaka, H. The melting temperature of proton-disordered hexagonal ice: A computer simulation of TIP4P model of water. *J. Chem. Phys.* **112**, 8534–8538 (2000).

- Jorgensen, W. L. *et al.* Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
- Steele, W. A. *Interaction of Gases with Solid Surfaces* (Pergamon, Oxford, 1974).
- Rapaport, D. C. *The Art of Molecular Dynamics Simulations* (Cambridge Univ. Press, Cambridge, 1997).

Acknowledgements

K.K. and H.T. are supported by the Japan Society for the Promotion of Science (JSPS), the Japan Ministry of Education, and Institute of Molecular Science. X.C.Z. is supported by the US National Science Foundation and Office of Naval Research, and by a JSPS fellowship.

Correspondence and requests for materials should be addressed to K.K. (e-mail: kenkoga@fukuoka-edu.ac.jp).

Efficient silicon light-emitting diodes

Martin A. Green*, Jianhua Zhao†, Aihua Wang†, Peter J. Reece‡ & Michael Gal‡

* Centre for Third Generation Photovoltaics; † Photovoltaics Special Research Centre; and ‡ School of Physics, University of New South Wales, Sydney, New South Wales 2052, Australia

Considerable effort is being expended on the development of efficient silicon light-emitting devices compatible with silicon-based integrated circuit technology¹. Although several approaches are being explored^{1–6}, all presently suffer from low emission efficiencies, with values in the 0.01–0.1% range regarded as high². Here we report a large increase in silicon light-emitting diode power conversion efficiency to values above 1% near room temperature—close to the values of representative direct bandgap emitters of a little more than a decade ago^{7,8}. Our devices are based on normally weak one- and two-phonon assisted sub-bandgap light-emission processes. Their design takes advantage of the reciprocity between light absorption and emission by maximizing absorption at relevant sub-bandgap wavelengths while reducing the scope for parasitic non-radiative recombination within the diode. Each feature individually is shown to improve the emission efficiency by a factor of ten, which accounts for the improvement by a factor of one hundred on the efficiency of baseline devices.

Planck's⁹ studies of light emission from bodies at temperature T resulted in the following expression for $\dot{n}$, the photon emission rate per unit area, energy and solid angle Ω :

$$\dot{n} dE d\Omega = \frac{2}{h^3 c^2} A(E) E^2 f_{pt} dE d\Omega \quad (1)$$

where E is the photon energy, c is the velocity of light, and h and k are the Planck and Boltzmann constants. $A(E)$ is the body's spectral absorptance, earlier shown by Kirchhoff¹⁰ to equal its spectral emissivity. f_{pt} is the equilibrium photon state occupation probability:

$$f_{pt} = 1/[\exp(E/kT) - 1] \quad (2)$$

More recently, Planck's theory has been generalized to describe light emission from semiconductor p–n junctions¹¹. In these devices, light arises from recombination between conduction band electrons and valence band holes, with distributions in each band described by different electrochemical potentials (quasi-Fermi energies). Rather than assigning zero chemical potential to photons, as in equation (2), this generalization assigns a chemical potential μ to the emitted light, equal to the difference between the electron and hole quasi-Fermi energies. The photon state occupation probability for photons in equilibrium with the electron–hole distribution

then becomes¹¹:

$$f_{pt} = 1/\{\exp[(E - \mu)/kT] - 1\} \quad (3)$$

The above generalization allows a simple formulation for light emission from high-quality diodes with carrier diffusion lengths larger than their thickness, low surface recombination rates and negligible series resistance. In this case, electron and hole quasi-Fermi energies are approximately constant throughout the diode, separated by energy qV , where V is applied voltage and q the electronic charge. For photon energies where absorption is predominantly band-to-band, normally those above the semiconductor bandgap, the original Planckian expression (equation (1)) again applies but with the equilibrium photon occupation probability becoming:

$$f_{pt1} = 1/\{\exp[(E - qV)/kT] - 1\} \approx f_{pt0} e^{qV/kT} \quad (4)$$

Kirchhoff's law also applies to devices meeting the previous conditions¹², even when the semiconductor involved has an indirect bandgap¹³. Light emission above the bandgap in high-quality p-n junctions can be regarded as Planckian radiation exponentially enhanced by diode voltage¹¹, as deduced earlier more intuitively by the following argument¹⁴.

A diode with zero applied voltage in thermal equilibrium at temperature T is in equilibrium with encompassing blackbody radiation at this temperature. At each wavelength above the bandgap, the diode absorbs a fraction of the incident blackbody radiation given by its spectral absorptance and emits the same fraction. The physical source of this emitted super-bandgap radiation is predominantly band-to-band recombination within the diode. Much of the light emitted during this recombination is re-absorbed, internally reflected and so on within the diode, with only a fraction reaching the exterior, balancing the incoming light. If a voltage is now applied to the diode, with the diode assumed to be high quality as defined above, the electron-hole concentration product will increase exponentially throughout the diode volume, exponentially increasing the radiative recombination rate throughout. As the same fraction of this is re-absorbed internally, reflected, and so on, as before, the emitted light increases exponentially by the same factor, giving the approximate form of equation (4). (Stimulated emission needs to be considered to derive the full expression.)

Parasitic internal absorption, such as by free carriers or at metal contacts, complicates this simple picture. Free carrier and contact absorption involves excitations within a single band and hence emission with zero chemical potential. Equilibrium photon occupation probabilities when processes with different chemical potentials are involved have been analysed recently in another context¹⁵. Assuming parasitic absorption can be modelled as occurring uniformly within the diode—a good assumption for weakly absorbed

light—these results can be extended to the present case. The Planckian result again applies with f_{pt} given by:

$$f_{pt2} = \frac{\alpha_{bb} e^{-(E-qV)/kT} + \alpha_{pa} e^{-E/kT}}{\alpha_{bb}[1 - e^{-(E-qV)/kT}] + \alpha_{pa}[1 - e^{-E/kT}]} \approx \frac{\alpha_{bb} f_{pt0}}{\alpha_{bb} + \alpha_{pa}} e^{qV/kT} \quad (5)$$

α_{bb} and α_{pa} are band-to-band and parasitic absorption coefficients. The approximate version of equation (5) also has been derived elsewhere¹⁶. The requirement for strong light emission for a given wavelength, direction and voltage, in a high-quality diode, is high absorptance for light incident from the opposite direction, but with low parasitic absorption.

A technique used in photovoltaics to increase absorptance is “trapping” of weakly absorbed light^{17,18} by taking advantage of small internal escape cones. For example, upon randomizing light direction once entering semiconductor of refractive index n , only a fraction $1/(2n^2)$ lies within the escape cone of any single surface. Randomization occurs with diffusely scattering (Lambertian) surface textures although regular surface structures can perform similarly or better for directional light¹⁹. Absorptance approximately equals in these cases:

$$A \approx (1 - R_f)(1 - t)/(1 - R_f t) \quad \text{where} \quad t \approx \exp(-4\alpha W) \quad (6)$$

t is the double-pass transmission, R_f is the front-surface reflectance, W is the device thickness and α is given by¹⁵:

$$\alpha = \alpha_{bb}\{1 - \exp[(qV - E)/kT]\} + \alpha_{pa}\{1 - \exp(-E/kT)\} \approx \alpha_{bb} + \alpha_{pa} \quad (7)$$

For weak absorption, trapping increases absorptance by a factor of up to $2n^2$, about 25 times that for typical semiconductors. Although light randomization has recently been applied to direct bandgap diodes to increase emission efficiency^{20,21}, the advantages for silicon are greater and are a major contributor to the improved performance reported here.

For high emission efficiency, the other factor that needs to be optimized is parasitic non-radiative recombination in the diode. Under forward bias, current density J approximately equals:

$$J = J_{o1} \exp(qV/kT) + J_{o2} \exp(qV/2kT) \quad (8)$$

The J_{o1} “unity ideality factor” component describes bulk radiative recombination as well as non-radiative recombination in those bulk, surface and contact regions where either electrons or holes form a strong majority. The J_{o2} component characterizes non-radiative recombination in regions where electron and hole concentrations are approximately equal, such as within junction transition regions, occasionally at surfaces, or throughout lightly doped

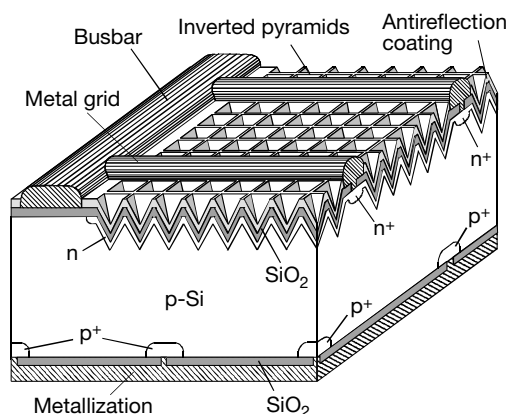


Figure 1 Schematic of high-efficiency silicon light-emitting diode.

regions. In silicon, where the J_{02} component can be relatively small, peak power conversion efficiency can be much higher than peak quantum efficiency, because the average energy of emitted photons can be much higher than qV near the peak. This effect is not observed to any significant extent in III–V diodes owing to relatively high J_{02} .

This background allows the advantages of the improved device shown in Fig. 1 over previously reported silicon light emitters to be explained. High absorptance (hence emissivity) is obtained using inverted pyramids on the top surface, formed by anisotropic etching to expose (111) equivalent crystallographic planes in the originally (100) surface. These pyramids not only reduce reflection but, more importantly, increase absorptance by trapping weakly absorbed light within the cell. Even though light is not immediately randomized upon entering the diode, it is randomized once it passes across the cell and back, because it has then been internally reflected by the pyramids many times. Light-trapping performance on the first few passes is similar to that of a

randomizing scheme, so the previous Lambertian theory describes absorptance well.

Additional features enhancing absorptance include the rear metal reflector, which ideally doubles the effective thickness, an anti-reflection coating that reduces R_f and a reasonable device thickness (0.45 mm). To reduce parasitic absorption, wafers are only moderately doped ($\sim 1.4 \times 10^{16} \text{ cm}^{-3}$) and dopant diffusions are either relatively light or confined to small volumes, reducing free carrier absorption (bulk and diffused areas contribute approximately equally to the estimated effective free carrier absorption coefficient of 0.1 cm^{-1} at emission wavelengths). Absorption in metal contacts and reflectors increases the parasitic absorption coefficient by an estimated 0.2 cm^{-1} , whereas α_{bb} drops from 3.5 cm^{-1} at silicon's bandgap (1,103 nm) to 10^{-3} cm^{-1} at 1,250 nm (ref. 22).

To minimize parasitic recombination, devices are fabricated using wafers prepared by the float-zone growth process, giving high minority-carrier recombination lifetimes; surfaces are passivated by high-quality thermal oxides; metal contact areas are kept

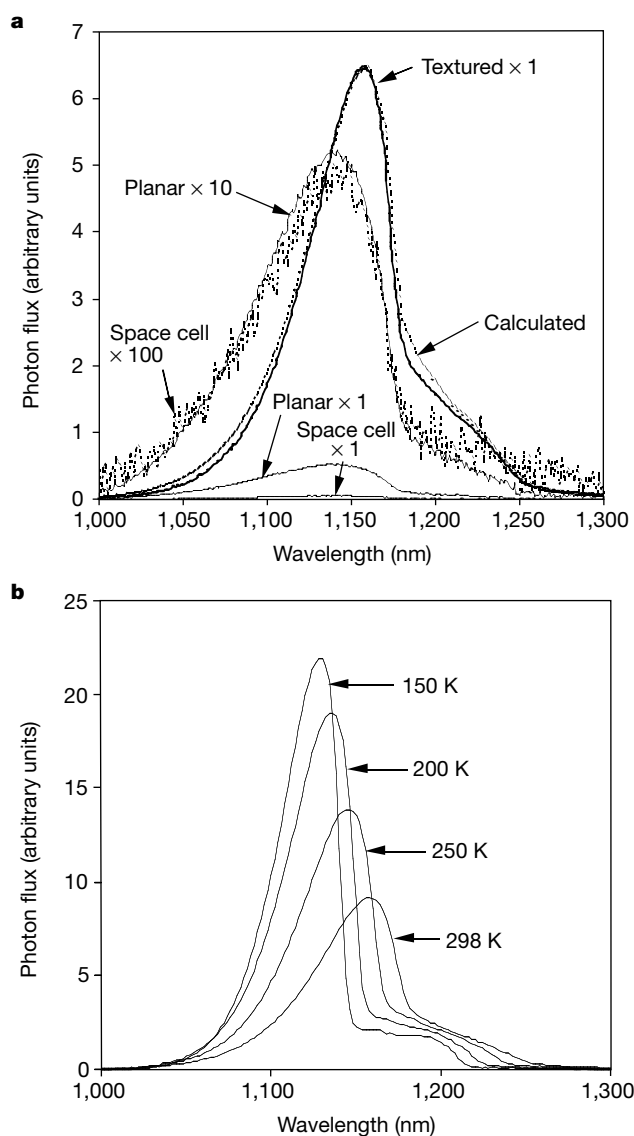


Figure 2 Electroluminescence spectra for various silicon diodes and temperatures. **a**, Spectra for textured, planar and baseline space cell diodes under 130 mA bias current at 298 K (diode area 4 cm^2). Calculated value assumes a rear reflectance of 96%. **b**, Temperature dependence of electroluminescence for the textured device under 100 mA bias.

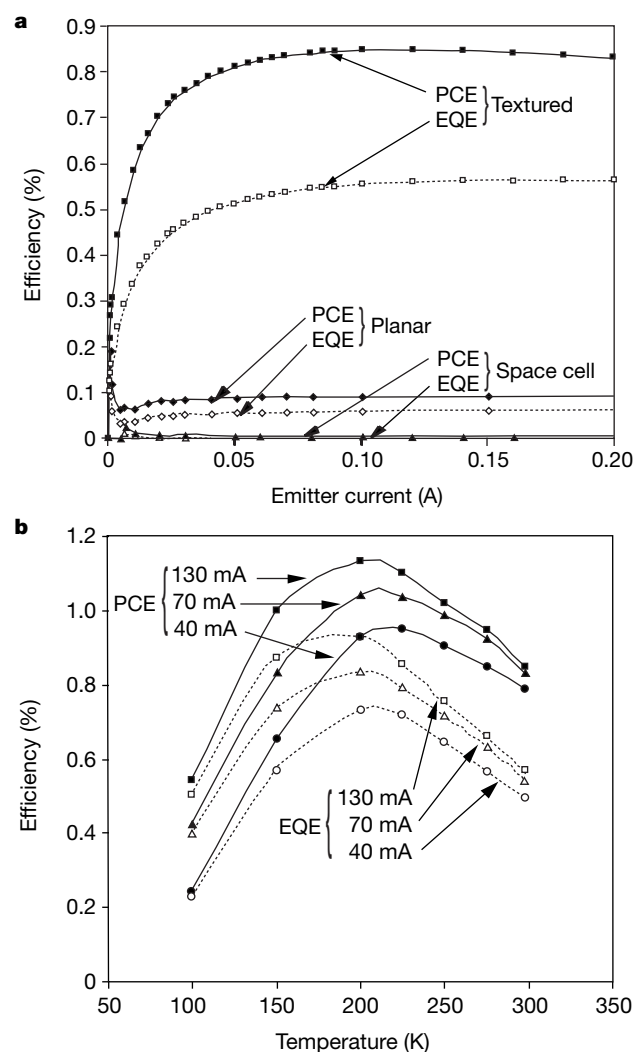


Figure 3 External quantum efficiency and power conversion efficiency. Filled symbols and solid lines indicate power conversion efficiency (PCE); open symbols and dotted lines indicate external quantum efficiency (EQE). **a**, Values at 294 K for the three diodes shown in Fig. 2a as a function of diode current (diode area 4 cm^2). Squares, textured diode; diamonds, planar diode; triangles, space cell diode. **b**, Temperature dependence of external quantum efficiency (EQE) and power conversion efficiency for the textured diode for three different diode currents: 130 mA (squares); 70 mA (triangles); 40 mA (circles).

small; and heavy doping is localized to contact areas. Such measures reduce J_{01} and J_{02} to low values with J_{01} typically 70 fA cm^{-2} at 300 K. Post-processing minority carrier lifetimes in the bulk of these devices are close to or above 1 ms (ref. 23), not far below radiative lifetimes estimated to be in the 7–14-ms range at the present doping level^{24,25}. Similar technology has produced the highest-ever silicon solar cell performance²³ (24.7% efficiency compared to 25.1% for the best GaAs device).

Wavelength dependence of perpendicular emission at 298 K is shown as the uppermost curve of Fig. 2a. Normalized outputs at diode currents in the 10 mA to 130 mA range tested are spectrally essentially identical. The spectral line shapes are well described by the previous theory, confirming emission via band-to-band processes. The inflection apparent at 1,180 nm corresponds to the transition between processes involving emission of one and two phonons, primarily the 58 meV transverse optical (TO) phonon and this plus the 64 meV Raman phonon, respectively. Processes involving three phonons (additional Raman phonon) appear at longer wavelengths. Each additional Raman phonon past this stage reduces emission intensity about 10 times (for example, emission at a wavelength of 1,550 nm involves five phonons, with intensity in present devices estimated as 1,000 times smaller than at 1,250 nm).

Figure 2a also shows emission on the same scale at the same current density for two other silicon diodes. The lowermost solid curve, barely distinguishable from the horizontal axis, shows emission from one of the presently most developed commercial silicon optoelectronic devices, a high-performance space solar cell. This device has a peak quantum efficiency of 0.004%, consistent with previous reports for silicon diodes^{26,27}. As shown by the far left curve (output magnified 100 times), the present textured device outperforms this baseline by about a hundred-fold.

The second-lowest curve of Fig. 2a is for a device of the same design as Fig. 1 but with a planar top surface, without pyramids. Compared to the baseline, the increased output arises primarily from reduced non-radiative recombination, showing a tenfold improvement. Compared to the textured device, the latter's higher absorptance at sub-bandgap wavelengths increases output by close to the theoretical value of 25 times at long wavelengths, but tenfold overall.

The emission spectra at various temperatures is shown in Fig. 2b. As temperature decreased to 150 K, emission intensity increased. The emission peak follows the temperature dependence of the silicon bandgap. At 150 K, two humps appear in the two-phonon emission region, investigated previously via photoluminescence²⁸. The longer wavelength feature is attributed to the TO phonon plus Raman; the other to a combination of phonon pairs including the transverse acoustic (TA) plus Raman and the TO plus intervalley phonons. The increasing emission as temperature decreases is attributed primarily to a difference in the activation energy of radiative and non-radiative processes. The latter equals the bandgap whereas the former is lower by the excitonic binding energy (15 meV). Below 150 K, emission efficiency decreases. This is attributed to increased non-radiative recombination that is a result of increased surface velocities at low temperatures, as deduced from low-temperature spectral-response measurements on similar diodes²⁹.

The external quantum conversion efficiencies of the three devices shown in Fig. 2a were measured at 294 K using a calibrated Ge detector and are shown in Fig. 3a as a function of diode current, along with power conversion efficiencies (edge emission, estimated as 10% of the total, excluded). Power conversion efficiency of the textured device peaks at close to 1%, over 100 times greater than that of the silicon baseline. The fall-off at low current is due to increasing non-radiative recombination, including junction recombination as per equation (8). A slight fall-off in external quantum conversion efficiency at higher current is also observed, although its

origins have not been conclusively identified as yet (the fall-off could be due to current crowding around the metal grid of Fig. 1 or increasing bulk non-radiative Auger recombination or excitonic screening as high injection is approached). The absolute measurements near 300 K allow the spectral results (as in Fig. 2b) to be converted to absolute measurements across the temperature range (as in Fig. 3b).

The efficiency could be further increased by increasing diode thickness, reducing free carrier concentrations, increasing rear mirror reflectance, or by removing the rear reflector and using the emission from both surfaces. Emission efficiency would be higher into material of higher refractive index than air, such as optical fibres or silicon waveguides. The high radiative lifetimes in silicon make efficient devices such as these relatively slow to respond, although integrated modulators may be preferred even for more quickly responding direct-gap devices³⁰. □

Received 11 May; accepted 26 July 2001.

1. Bell, P. Let there be light. *Nature* **409**, 974–976 (2001).
2. Ng, W. L. *et al.* An efficient room-temperature silicon-based light-emitting diode. *Nature* **410**, 192–194 (2001).
3. Vescan, L. & Stoica, T. Room-temperature SiGe light-emitting diodes. *J. Luminescence* **80**, 485–489 (1999).
4. Leong, D., Harry, M., Reeson, K. J. & Homewood, K. P. A silicon/iron disilicide light-emitting diode operating at a wavelength of 1.5 μm . *Nature* **387**, 686–688 (1997).
5. Hirschman, K. D., Tybekov, L., Duttaputra, S. P. & Fauchet, P. M. Silicon-based visible light-emitting devices integrated into microelectronic circuits. *Nature* **384**, 338–341 (1996).
6. Zheng, B. *et al.* Room-temperature sharp line electroluminescence at $\lambda = 1.54 \mu\text{m}$ from an erbium-doped, silicon light-emitting diode. *Appl. Phys. Lett.* **64**, 2842–2844 (1994).
7. Schnitzer, I., Yablonovitch, E., Caneau, C. & Gmitter, T. J. Ultrahigh spontaneous emission quantum efficiency, 99.7% internally and 72% externally, from AlGaAs/GaAs/AlGaAs double heterostructures. *Appl. Phys. Lett.* **62**, 131–133 (1993).
8. Meyer, M. "Craford's Law" and the evolution of the LED industry. *Compound Semicond.* **6**(2), 26–30 (2000).
9. Planck, M. *The Theory of Heat Radiation* (Dover, New York, 1959); translation of *Vorlesungen über die Theorie der Wärmestrahlung* (Leipzig, Barth, 1913).
10. Kirchhoff, G. R. Über das Verhältnis zwischen Emissionsvermögen und dem Absorptionsvermögen der Körper für Wärme und Licht. *Ann. D. Phys.* **109**, 275–301 (1860) (in German).
11. Würfel, P. The chemical potential of radiation. *J. Phys. C* **15**, 3967–3985 (1982).
12. Araújo, G. L. & Marti, A. Absolute limiting efficiencies for photovoltaic energy conversion. *Solar Energy Mater. Solar Cells* **33**, 213–240 (1994).
13. Würfel, P., Finkbeiner, S. & Daub, E. Generalised Planck's radiation law for luminescence via indirect transitions. *Appl. Phys. A* **60**, 67–70 (1995).
14. Shockley, W. & Queisser, H. J. Detailed balance limit of efficiency of p–n junction solar cells. *J. Appl. Phys.* **32**, 510–519 (1961).
15. Luque, A. & Marti, A. Increasing the efficiency of ideal solar cells by photon induced transitions at intermediate levels. *Phys. Rev. Lett.* **78**, 5014–5017 (1997).
16. Trupke, T., Daub, E. & Würfel, P. Absorptivity of silicon solar cells obtained from luminescence. *Solar Energy Mater. Solar Cells* **53**, 103–114 (1998).
17. Goetzberger, A. Optical confinement in thin Si solar cells by diffuse back reflectors. *Conf. Record, 15th IEEE Photovoltaic Specialists Conf.* Orlando, 867–870 (IEEE, New York, 1981).
18. Yablonovitch, E. Statistical ray optics. *J. Opt. Soc. Am.* **72**, 899–907 (1982).
19. Campbell, P. & Green, M. A. Light trapping properties of pyramidally textured surfaces. *J. Appl. Phys.* **62**, 243–249 (1987).
20. Schnitzer, I. & Yablonovitch, E. 30% external quantum efficiency from surface textured, thin-film light-emitting diodes. *Appl. Phys. Lett.* **63**, 2174–2176 (1993).
21. Windisch, R. *et al.* 40% efficient thin-film surface-textured light-emitting diodes by optimisation of natural lithography. *IEEE Trans. Electron Dev.* **47**, 1492–1498 (2000).
22. Green, M. A. & Keevers, M. J. Optical properties of intrinsic silicon at 300K. *Prog. Photovoltaics* **3**, 189–192 (1995).
23. Zhao, J., Wang, A. & Green, M. A. 24.5% efficiency PERT silicon solar cells on SEH MCZ substrates and cell performance on other SEH CZ and FZ substrates. *Solar Energy Mater. Solar Cells* **66**, 27–36 (2001).
24. Schlagenotto, H., Maeder, H. & Gerlach, W. Temperature dependence of the radiative recombination coefficient in silicon. *Phys. Status Solidi; (A) Applied Research* **21**, 357–367 (1974).
25. Wasserrab, Th. Beitrag zur Berechnung der strahlenden Rekombinations-wahrscheinlichkeit B von eigenleitendem Silicium. *Z. Naturforsch; Part A* **33**, 1097–1098 (1978) (in German).
26. Ong, T.-C., Terrill, K. W., Tam, S. & Hu, C. Photon generation in forward-biased silicon p–n junctions. *IEEE Electron Dev. Lett.* **4**, 460–462 (1983).
27. Kramer, J. *et al.* Light-emitting devices in industrial CMOS technology. *Sensors Actuators A* **37–38**, 527–533 (1993).
28. Vouk, M. A. & Lightowlers, E. C. Two-phonon assisted free exciton recombination radiation from intrinsic silicon. *J. Phys. C; Solid State Physics* **10**, 3689–3699 (1977).
29. Neisser, A. *Spectral Response Measurements on Silicon Solar Cells in the Range of 1 eV to 5 eV Photon Energy at Different Temperatures*. Thesis, Technische Universität Berlin (1998).
30. Thomas, G. A., Ackerman, D. A., Prucnal, P. R. & Cooper, S. L. Physics in the whirlwind of optical communications. *Phys. Today* **53**(9), 30–36 (2000).

Correspondence and requests for materials should be addressed to M.A.G. (e-mail: m.green@unsw.edu.au).

Intensified deep Pacific inflow and ventilation in Pleistocene glacial times

Ian R. Hall⁺, I. Nicholas McCave[†], Nicholas J. Shackleton[†],
Graham P. Weedon[‡] & Sara E. Harris[§]

⁺ Department of Earth Sciences, Cardiff University, PO Box 914, Cardiff CF10 3YE, UK

[†] Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK

[‡] Department of Environment, Geography and Geology, University of Luton, Park Square, Luton LU1 3JU, UK

[§] Sea Education Association, PO Box 6, Woods Hole, Massachusetts 02543, USA

The production of cold, deep waters in the Southern Ocean is an important factor in the Earth's heat budget¹. The supply of deep water to the Pacific Ocean is presently dominated by a single source, the deep western boundary current east of New Zealand. Here we use sediment records deposited under the influence of this deep western boundary current to reconstruct deep-water properties and speed changes during the Pleistocene epoch. In physical and isotope proxies we find evidence for intensified deep Pacific Ocean inflow and ventilation during the glacial periods of the past 1.2 million years. The changes in throughflow may be directly related to an increased production of Antarctic Bottom Water during glacial times. Possible causes for such an increased bottom-water production include increasing wind strengths in the Southern Ocean or an increase in annual sea-ice formation, leaving dense water after brine rejection and thereby enhancing deep convection. We infer also that the global thermohaline circulation was perturbed significantly during the mid-Pleistocene climate transition between 0.86 and 0.45 million years ago.

The main deep-water masses in the oceans originate in the North Atlantic Ocean and around Antarctica, where several sources make up Antarctic Bottom Water (AABW, in the broad sense). The long-term behaviour of the sources of North Atlantic Deep Water (NADW) has been investigated in detail through analysis of deep-sea cores at many sites (see, for example, ref. 2). There is physical evidence that at times, particularly during glacials, the inflow of NADW was slower than in the Holocene epoch³. The equally important history of AABW is relatively unstudied. The influx of AABW to the oceans is principally via the southwest Atlantic, east and west Indian, and southwest Pacific oceans. Here we present records (for the past 1.2 Myr) of glacial–interglacial (G–I) variability of the strength and water-mass properties of the inflow to the southwest Pacific; this inflow is represented by the deep western boundary current (DWBC) flowing east of New Zealand (Fig. 1), which, in terms of water flux, is the largest DWBC in the world ocean.

Ocean Drilling Program (ODP) Site 1123 (41° 47.2' S, 171° 29.9' W), located at 3,290 m water depth⁴ (Fig. 1), lies beneath the southwest Pacific DWBC, whose upper boundary is conventionally placed at ~2,000 m, corresponding to the level of no motion⁵. The core sampled a major sediment drift that is deposited as the DWBC decelerates after passing through Valerie passage, and so is well placed to monitor the deep Pacific inflow⁴. Samples were analysed every 5–10 cm (or ~1.2–2.5 kyr), spanning a 47-m section of core (Fig. 2). The timescale for Site 1123 was constructed from the benthic $\delta^{18}\text{O}$ record (Fig. 2a), using an ice-volume model⁶ as the tuning target: we generated the latter using July insolation values at 65° N (ref. 7). The segment of core that we analysed records the past 1.2 Myr, with sedimentation rates between 2.1 and 5.1 cm kyr⁻¹. Ages for the magnetic polarity reversals based on our timescale are in good agreement with previous estimates⁸.

We used two proxies to examine the G–I variations in the thermohaline circulation (Fig. 2b, c). First, the sortable silt mean size (SS): this is defined as the mean grain size of the 10–63 μm terrigenous fraction, with biogenic carbonate and opal removed. It is a sedimentological parameter where coarser mean size reflects stronger near-bottom flow, through selective deposition and winnowing⁹. In regions of high eddy kinetic energy (K_E), there is some potential uncertainty in assessing the significance of the parameter for mean flow⁹. However, Site 1123 lies at a location where estimates of benthic K_E are < 1 to 2 cm² s⁻², well north of the Antarctic Circumpolar Current (ACC), with its high K_E (up to 80 cm² s⁻²) and benthic storms¹⁰. Thus mean flow is expected to dominate the SS signal. The similarity of behaviour of this parameter at separated sites in the Atlantic deep circulation system has strengthened confidence in its ability to reveal regional flow changes without local short-term interference¹¹. Second, we used the carbon isotopic ($\delta^{13}\text{C}$) values of benthic foraminifera, which are related to the nutrient concentration of the waters bathing the site. In the North Atlantic, where the residence time for deep waters is ~200 years, $\delta^{13}\text{C}$ mainly indicates the relative proportions of nutrient-depleted NADW (high $\delta^{13}\text{C}$) and nutrient-enriched southern source water (low $\delta^{13}\text{C}$)²; but the $\delta^{13}\text{C}$ signal also evolves (or ages) along a flow path, because carbon with a low $\delta^{13}\text{C}$ signature is continuously released to bottom water by the oxidation of organic matter raining in from the surface. Owing to a longer deep-water residence time, the ageing effects are much more important in the Pacific than in other oceans².

The proxy flow-speed data in Fig. 2c clearly show faster flow in glacial periods and slower flow during interglacials (Fig. 3a). Multi-taper spectral analysis of the SS record shows significant spectral

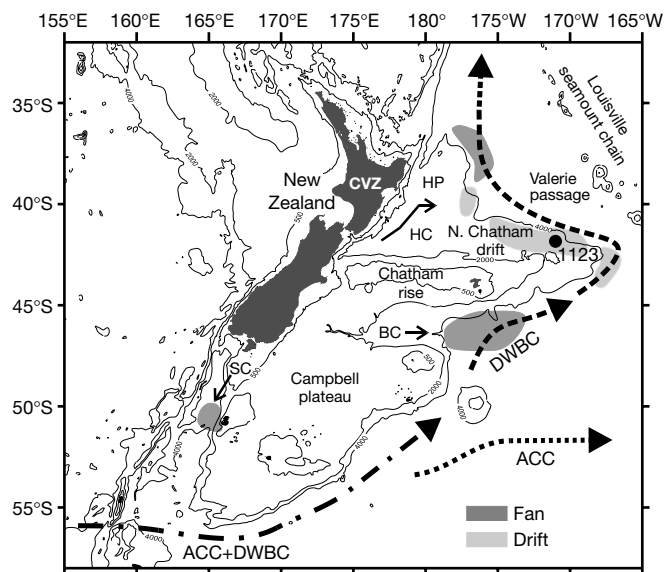


Figure 1 Location of study area. Shown are a generalized abyssal circulation scheme, sediment fans and drifts, which together make up the major elements of the eastern New Zealand oceanic sedimentary system (ENZOSS; after ref. 28). Contours are at depths of 500, 2,000 and 4,000 m. Sediment is supplied down the Solander (SC), Bounty (BC) and Hikurangi (HC) channels to abyssal depths and into the path of the deep western boundary current (DWBC), which transports the material along the margin of Chatham Rise/Hikurangi Plateau, depositing sediment drifts where the flow slows down. The location of Site 1123 on north Chatham drift is shown. Other sources of terrigenous material for the region are dust carried by the westerly winds from New Zealand and possibly South Australia, and fallout of rhyolitic and andesitic ash from explosive volcanism in New Zealand. ACC, Antarctic Circumpolar Current; CVZ, central volcanic zone; HP, Hikurangi Plateau.

peaks at each orbital frequency; these peaks are coherent with both isotope records at confidence limits of 98% (for orbital periods of 100 and 41 kyr) and 90% (for the 23-kyr orbital period). There are zero phase lags between $\overline{SS}$ and benthic $\delta^{18}O$ at the 100-kyr period, but a small lag of 2.1 ± 1.4 kyr at 41 kyr (see Fig. 3c, and Supplementary Information). These results resolve the paradox that, although the DWBC has a very large flux, it appears at present to be sluggish around the New Zealand margin (shown by current meters, nepheloid layers, and bottom photographs), yet there are very large scours around volcanic pinnacles on the sea bed¹². Our results suggest that the scours probably formed under strong glacial bottom-flows. Further downstream, these enhanced flows may have produced the circulation changes that drove the glacial increase in sediment focussing recorded in the central equatorial Pacific over the past 300 kyr (ref. 13).

The SS time series also shows three periods of differing mean flow speeds of the DWBC; first, a period of moderately high flow before termination of glacial marine isotope stage (MIS) 22 (the first large 100-kyr cycle seen in $\delta^{18}O$, centred around 900 kyr ago, Fig. 2a); second, a transition period of lower flow speed from 870 to 450 kyr

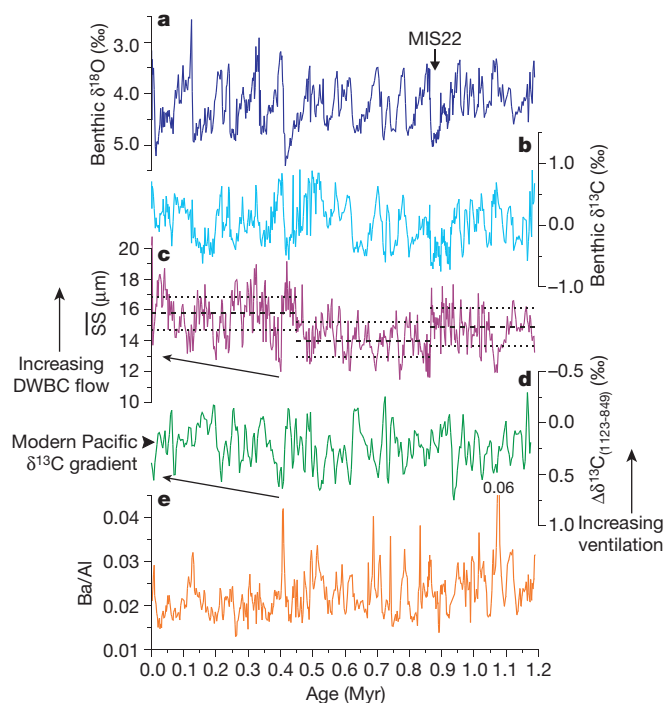


Figure 2 Summary of Site 1123 data. Shown are the records of: **a**, benthic $\delta^{18}O$; **b**, benthic $\delta^{13}C$; **c**, $\overline{SS}$ (sortable silt mean grain size; see text); **d**, $\delta^{13}C$ gradient for the southwest Pacific-east equatorial Pacific ($\Delta\delta^{13}C_{(1123-849)}$; note reversed scale); **e**, bulk Ba/Al ratios plotted to age. Isotopic measurements were made on the benthic foraminifera genus *Uvigerina*. All isotopic determinations were performed using a VG Sira Series II mass spectrometer and Multiprep system. Isotope values are reported relative to Pee Dee Belemnite using the standard NBS-19. Analytical precision is $\pm 0.08\text{‰}$ for $\delta^{18}O$ and $\pm 0.06\text{‰}$ for $\delta^{13}C$ measurements. To account for the 'vital effect', *Uvigerina* $\delta^{13}C$ was adjusted by $+0.9\text{‰}$. Grain size distributions of the fine terrigenous fraction were determined using a Micromeritics SediGraph 5100 whose analytical precision is between $\pm 2\text{--}13\%$ for $\overline{SS}$ ²⁹. The dashed and dotted lines indicate the mean and standard errors respectively for the periods 0–0.45 Myr ago, 0.45–0.87 Myr ago, and 0.87–1.19 Myr ago. For the estimates of $\Delta\delta^{13}C_{(1123-849)}$, benthic $\delta^{18}O$ records for Site 849¹⁸ were correlated to the Site 1123 timescale and both sites were converted to equal time intervals of 3.5 kyr by gaussian interpolation. The record for Site 849 is from analysis only of *Cibicides*. The elemental composition of the bulk sediment was determined using a Perkin Elmer Plasma 40 Philips PV8060 inductively coupled plasma–atomic emission spectrometer. MIS, marine isotope stage.

ago; and last, the higher mean flow up to the present (Fig. 2c). The means of these periods are significantly different ($P < 0.001$). The middle period corresponds to the mid-Pleistocene climate transition; this marks the period of change in the dominant response of the Earth's climate from orbital obliquity to eccentricity forcing¹⁴. This pattern is consistent with the view that the mid-Pleistocene

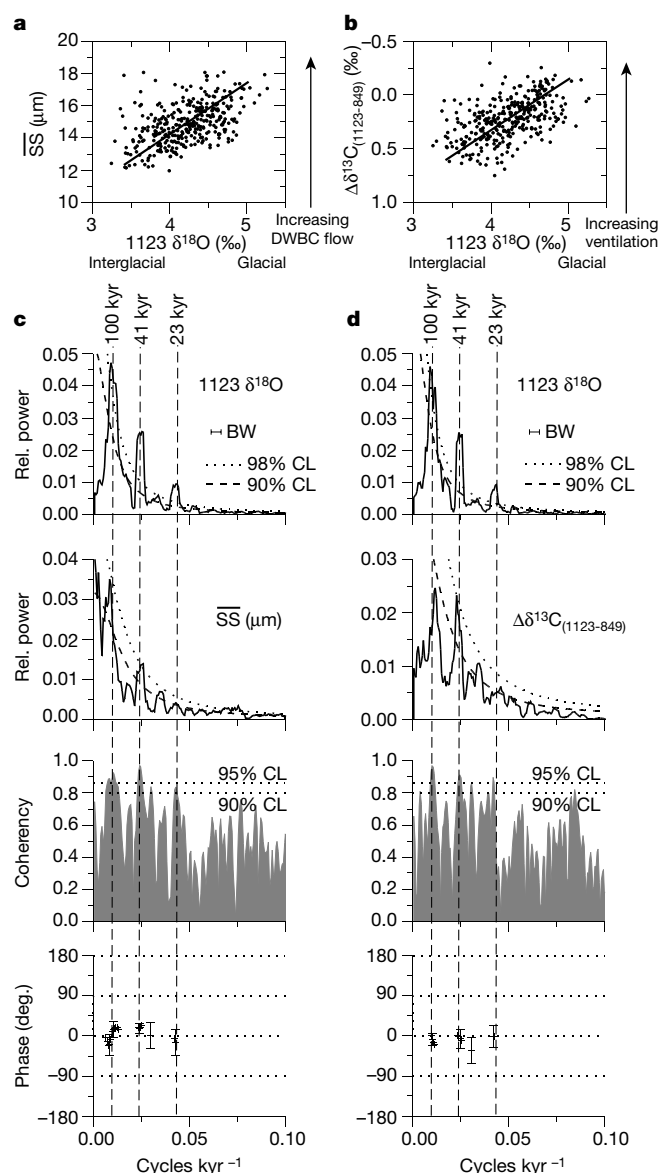


Figure 3 Summary of Pacific ventilation indicators. **a**, Site 1123 $\delta^{18}O$ versus $\overline{SS}$; **b**, Site 1123 $\delta^{18}O$ versus $\Delta\delta^{13}C_{(1123-849)}$ (note reversed scale); **c**, spectra and cross-spectra of the variables in **a**; **d**, spectra and cross-spectra of the variables in **b** (see Supplementary Information for details of spectral analysis, coherence and phase). For **a** and **b**, the fitted lines represent reduced major axis regression of the data. Power spectra were calculated using the multi-taper method³⁰ with seven data-tapers and spectral confidence levels located using the robust AR(1) modelling of median-smoothed spectra³¹. Cross-spectra used the algorithms of Bloomfield³². For the confidence levels of the power- and coherence-spectra and the confidence intervals of the phase spectra, we allowed for the effect of zero-padding³². Before analysis, data were gaussian interpolated to a common time interval of 3 kyr for $\overline{SS}$ and $\delta^{18}O$ in **a**, and 3.5 kyr for $\Delta\delta^{13}C$ gradient data and $\delta^{18}O$ in **b**. Coherency indicates the correlation between the pair of variables as a function of frequency. The confidence levels for coherency are 90% = 0.801, 95% = 0.859. The phase indicates the relative timing of variations for the pairs of variables. Positive phase indicates that the first variable leads the second and vice versa. BW, bandwidth; CL, confidence level.

climate transition occurred over several hundred kyr rather than as a brief event¹⁴. It indicates that the vigour of a major component of the deep thermohaline circulation was involved in this change. Over the most recent period, the strength of the flows in the interglacials have clearly increased, a long-term trend that appears to have been an important feature of the past 0.9 Myr.

The benthic $\delta^{13}\text{C}$ record is in phase with $\delta^{18}\text{O}$ (that is, light $\delta^{13}\text{C}$ is associated with heavy $\delta^{18}\text{O}$) at major orbital frequencies, and has a range from nutrient-enriched -0.6‰ in glacial to depleted values of $+0.8\text{‰}$ during interglacials (Fig. 2b). In the strongly advective regime of a sediment drift under a major DWBC, the effects of local productivity on the $\delta^{13}\text{C}$ of benthic foraminifera are thought to be slight¹⁵. Certainly at Site 1123 the organic-carbon content of the sediments is relatively low ($<0.5\%$), sulphate reduction is minor in the upper 47 m (ref. 4), and mineral magnetic signatures show no evidence of anoxic diagenesis¹⁶. Additionally, we note that the time series of the Ba/Al ratio, a tracer often taken to be indicative of productivity¹⁷, is quite different to that for $\delta^{13}\text{C}$ (Fig. 2e). Together, these suggest that the benthic $\delta^{13}\text{C}$ signal at Site 1123 is not primarily determined by local productivity but rather by the hydrographic properties of the water mass.

Deep-water $\delta^{13}\text{C}$ spatial gradients have shed light on ageing trends of water masses². Site 1123 is ideally located for assessment of $\delta^{13}\text{C}$ gradients within the Pacific, as it represents the source of most of the deep water. This deep water travels across the Pacific, progressively accumulating nutrients and depleted $\delta^{13}\text{C}$ through the remineralization of sinking organic material as well as mixing with the overlying waters. In Fig. 2d (note reversed axis) we show the difference in $\delta^{13}\text{C}$ between Site 1123 and Site 849 from the deep (3,850 m) eastern equatorial Pacific¹⁸ (this difference is termed $\Delta\delta^{13}\text{C}_{(1123-849)}$). The Pacific $\Delta\delta^{13}\text{C}_{(1123-849)}$ gradient varies by over 0.5‰ , with larger gradients associated with interglacial periods (Fig. 3b). The spectrum of $\Delta\delta^{13}\text{C}_{(1123-849)}$ shows strong obliquity power ($>98\%$) and is coherent with $\delta^{18}\text{O}$ ($>98\%$) with the records in phase (Fig. 3d; see also Supplementary Information). A clear relationship can also be seen between $\Delta\delta^{13}\text{C}_{(1123-849)}$ and $\overline{\text{SS}}$, with periods of reduced ventilation inferred from the isotopes (high $\Delta\delta^{13}\text{C}_{(1123-849)}$) associated with reduced DWBC flow speeds (physical ventilation). These records are $>90\%$ coherent with zero phase lags at the 100- and 41-kyr periods. An important related feature of the records is the very large G–I variation of about 1.4‰ in the $\delta^{13}\text{C}$ of the water emanating from the ACC and flowing into the Pacific (a part of which is attributable to changes in the global carbon reservoir). Today the Pacific water mass $\delta^{13}\text{C}$ gradient is only about 0.2‰ (ref. 19), so that the observed variability in the $\Delta\delta^{13}\text{C}_{(1123-849)}$ implies that during many interglacials (including the early Holocene) the gradient was increased compared to today. Indeed, we note a decreasing interglacial $\Delta\delta^{13}\text{C}_{(1123-849)}$ gradient that is consistent with the long-term increase in the interglacial flow speed of the DWBC observed in the $\overline{\text{SS}}$. On the other hand, the fact that at many glacial extremes the $\delta^{13}\text{C}$ values of Site 1123 are even lighter than at Site 849 (that is, negative $\Delta\delta^{13}\text{C}_{(1123-849)}$) serves as a reminder of the complexities in this proxy. Nevertheless this is (to our knowledge) the first time that changes in the Pacific $\delta^{13}\text{C}$ gradient have been explicitly linked to changes in the rate of deep-water input (the relationship between upwelling velocity, oxygen utilization and $\delta^{13}\text{C}$ in the eastern Pacific has been modelled²⁰). Our data require a substantial input of nutrient-depleted NADW to the Southern Ocean during interglacials in addition to the reduced flux of lower-NADW in glacial²¹.

The data presented here are internally consistent in requiring greater vigour of glacial thermohaline circulation. Radiochemical records also suggest similar, or slightly enhanced, circulation during the last glacial period compared to the modern ocean²². Our results may appear to conflict with the notion of weak circulation based on the apparent greater age (derived from radiocarbon^{23,24} measurements) of the glacial Pacific water mass. This is not necessarily so,

because the 'age' of the deep water is intimately related to the ^{14}C content of the newly-formed bottom water which in turn depends on ^{14}C in the source water. It is known that in the North Atlantic at the Last Glacial Maximum the surface water 'age' was much higher than its present 400 years (ref. 25), and it may also be the case that in glacial times AABW formed with a very low ^{14}C content. Indeed, data from the southwest Pacific²⁴ suggest that this was the case.

It has been proposed that the ocean's deep overturning circulation is largely dictated by the magnitude of Ekman transport driven by winds at the latitude of Drake Passage²⁶. On the basis of modelling studies, Rahmstorf and England²⁷ concluded that such a mechanism has only a moderate influence on the formation of NADW ($\sim 75\%$ of NADW in their model was generated by thermohaline forcing alone). However, they found that the production of AABW depends strongly on the winds over the Southern Ocean. While this offers a possible control for the deep Pacific inflow it is not straightforward as, in their model, only a reduction in wind stress from its present level causes a collapse in Pacific bottom-water inflow. Increase in wind stress beyond present levels does not appear to show the required increase in flow. Another probable mechanism leading to an increase in the formation rate of AABW during glacial periods is greater annual production of sea ice, leading to water densification through brine rejection.

Our data show that surprises remain in the operation of the global thermohaline 'conveyor belt' circulation. The present work demonstrates G–I variability in the strength and water-mass properties of the inflow to the southwest Pacific. This interpretation is consistent with greater DWBC flow and greater Pacific ventilation being persistent features of the glacial periods of the past 1.2 Myr, and directly related to the formation rate of AABW. Further modelling studies on the formation of bottom waters of Antarctic origin, residence time in the ACC, and the deep Pacific circulation under glacial conditions are clearly needed. □

Received 31 January; accepted 10 July 2001.

- Clarke, A., Church, J. & Gould, J. in *Ocean Circulation and Climate—Observing and Modelling the Global Ocean* (eds Siedler, G. et al.) 11–30 (Academic, San Diego, 2001).
- Raymo, M. E., Oppo, D. W. & Curry, W. The mid-Pleistocene climate transition: A deep sea carbon isotopic perspective. *Paleoceanography* **12**, 546–559 (1997).
- McCave, I. N., Manighetti, B. & Beveridge, N. A. S. Circulation of the glacial North Atlantic inferred from grain-size measurements. *Nature* **374**, 149–151 (1995).
- Shipboard Scientific Party. Site 1123: North Chatham Drift—a 20 Ma record of the Pacific Deep Western Boundary Current. *Proc. ODP Init. Rep.* [CD-ROM] Leg 181, 1–184 (1999).
- Warren, B. A. Transpacific hydrographic sections at 43°S and 28°S : The SCORPIO Expedition—II Deep water. *Deep-Sea Res.* **20**, 9–38 (1973).
- Imbrie, J. & Imbrie, J. Z. Modeling the climatic response to orbital variations. *Science* **207**, 943–953 (1980).
- Berger, A. & Loutre, M. F. Insolation values for climate of the last 10 million years. *Quat. Sci. Rev.* **10**, 297–317 (1991).
- Bassinot, F. C. et al. The astronomical theory of climate and the age of the Brunhes-Matuyama magnetic reversal. *Earth Planet. Sci. Lett.* **126**, 91–108 (1994).
- McCave, I. N., Manighetti, B. & Robinson, S. G. Sortable silt and fine sediment size/composition slicing: Parameters for palaeocurrent speed and palaeoceanography. *Paleoceanography* **10**, 593–610 (1995).
- Carter, L. & Wilkin, J. Abyssal circulation around New Zealand—a comparison between observations and a global circulation model. *Mar. Geol.* **159**, 221–239 (1999).
- Hall, I. R., McCave, I. N., Chapman, M. R. & Shackleton, N. J. Coherent deep flow variation in the Iceland and American basins during the last interglacial. *Earth Planet. Sci. Lett.* **164**, 15–21 (1998).
- McCave, I. N. & Carter, L. Recent sedimentation beneath the Deep Western Boundary Current off Northern New Zealand. *Deep-Sea Res.* **44**, 1203–1237 (1997).
- Marcantonio, F. et al. Sediment focusing in the central equatorial Pacific Ocean. *Paleoceanography* **16**, 260–267 (2001).
- Jansen, J. H. F., Kuijpers, A. & Troelstra, S. R. A mid-Brunhes climatic event—long-term changes in global atmosphere and ocean circulation. *Science* **232**, 619–622 (1986).
- Mackensen, A. et al. The $\delta^{13}\text{C}$ in benthic foraminiferal tests of *Fontbotia wuellerstorfi* (Schwager) relative to the $\delta^{13}\text{C}$ of dissolved inorganic carbon in Southern Ocean deep water: Implications for glacial ocean circulation models. *Paleoceanography* **8**, 587–610 (1993).
- Lean, C. M. B. & McCave, I. N. Glacial to interglacial mineral magnetic and palaeoceanographic changes at Chatham Rise, SW Pacific Ocean. *Earth Planet. Sci. Lett.* **163**, 247–260 (1998).
- Dymond, J., Suess, E. & Lyle, M. Barium in deep-sea sediment: a geochemical proxy for paleoproductivity. *Paleoceanography* **7**, 163–182 (1992).
- Mix, A. C. et al. Benthic foraminifer stable isotope record from Site 849 (0–5 Ma): Local and global climate changes. *Proc. ODP Sci. Res.* **138**, 371–412 (1995).
- Kroopnick, P. M. The distribution of ^{13}C of ECO_2 in the world oceans. *Deep-Sea Res.* **32**, 57–84 (1985).

20. Kroopnick, P. M. The dissolved O_2 - CO_2 - ^{13}C system in the eastern equatorial Pacific. *Deep-Sea Res.* **21**, 211–227 (1974).
21. Charles, C. D. & Fairbanks, R. G. Evidence from Southern Ocean sediments for the effect of North Atlantic deep-water flux on climate. *Nature* **355**, 416–419 (1992).
22. Yu, E.-F., Francois, R. & Bacon, M. P. Similar rates of modern and last-glacial ocean thermohaline circulation inferred from radiochemical data. *Nature* **379**, 689–694 (1996).
23. Adkins, J. F. & Boyle, E. A. Changing atmospheric $\delta^{14}C$ and the record of deep water paleoventilation ages. *Paleoceanography* **12**, 337–344 (1997).
24. Sikes, E. L. *et al.* Old radiocarbon ages in the southwest Pacific Ocean during the last glacial period and deglaciation. *Nature* **405**, 555–559 (2000).
25. Bard, E. *et al.* The North-Atlantic atmosphere-sea surface ^{14}C gradient during the Younger Dryas climatic event. *Earth Planet. Sci. Lett.* **126**, 275–287 (1994).
26. Toggweiler, J. R. & Samuels, B. in *The Global Carbon Cycle* (ed. Heimann, H.) 333–366 (Springer, Berlin, 1993).
27. Rahmstorf, S. & England, M. H. Influence of southern hemisphere winds on North Atlantic Deep Water flow. *J. Phys. Oceanogr.* **27**, 2040–2054 (1997).
28. Carter, L., Carter, R. M., McCave, I. N. & Gamble, J. Regional sediment recycling in the abyssal S.W. Pacific Ocean. *Geology* **24**, 735–738 (1996).
29. Bianchi, G. G., Hall, I. R., McCave, I. N. & Joseph, L. Measurement of the sortable silt current speed proxy using the Sedigraph 5100 and Coulter Multisizer II: precision and accuracy. *Sedimentology* **46**, 1001–1014 (1999).
30. Pardo-Igúzquiza, E., Chica-Olmo, M. & Rodríguez-Tovar, J. CYSTRAT: a computer program for spectral analysis of stratigraphic successions. *Comput. Geosci.* **20**, 511–584 (1994).
31. Mann, M. E. & Lees, J. Robust estimation of background noise and signal detection in climatic time series. *Clim. Change* **33**, 409–445 (1996).
32. Bloomfield, P. *Fourier Analysis of Time Series: an Introduction* (Wiley, London, 1976).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank A. Fulop, M. Hall, L. Tiegang, P. Knutz and N. Walsh for experimental assistance, and the captain, crew and shipboard scientific party of Leg 181 for their help. Samples were provided by the Ocean Drilling Program (ODP), with the support of the Natural Environment Research Council and the UK ODP.

Correspondence and requests for materials should be addressed to I.R.H. (e-mail: Hall@cardiff.ac.uk).

Tectonic contraction across Los Angeles after removal of groundwater pumping effects

Gerald W. Bawden*, Wayne Thatcher*, Ross S. Stein*, Ken W. Hudnut† & Gilles Peltzer‡

* United States Geological Survey, Menlo Park, California 94025, USA

† United States Geological Survey, Pasadena, California 91106, USA

‡ University of California, Los Angeles, California 90095, USA

After the 1987 Whittier Narrows¹ and 1994 Northridge² earthquakes revealed that blind thrust faults represent a significant threat to metropolitan Los Angeles³, a network of 250 continuously recording global positioning system (GPS) stations^{4,5} was deployed to monitor displacements associated with deep slip on both blind and surface faults. Here we augment this GPS data with interferometric synthetic aperture radar imagery to take into account the deformation associated with groundwater pumping and strike-slip faulting. After removing these non-tectonic signals, we are left with 4.4 mm yr⁻¹ of uniaxial contraction across the Los Angeles basin, oriented N 36° E (perpendicular to the major strike-slip faults in the area). This indicates that the contraction is primarily accommodated on thrust faults⁶ rather than on the northeast-trending strike-slip faults. We have found that widespread groundwater and oil pumping obscures and in some cases mimics the tectonic signals expected from the blind thrust faults. In the 40-km-long Santa Ana basin, groundwater withdrawal and re-injection produces 12 mm yr⁻¹ of long-term subsidence, accom-

panied by an unprecedented seasonal oscillation of 55 mm in the vertical direction and 7 mm horizontally.

Anthropogenic deformation is evident from a series of 1997–1999 interferograms, which reveal ongoing subsidence of the Wilmington oil field near Long Beach⁷, and seasonal uplift and subsidence of the Santa Ana basin, the primary groundwater source for Orange County (Fig. 1). The 5-km-wide Wilmington oil field (W in Fig. 1b) undergoes episodic subsidence (up to 30 mm over 175 days). The 20 km × 40 km Santa Ana basin displays seasonal fluctuations of 50 mm of basin uplift during the late fall through to mid-spring (Fig. 1b and c), and 60 mm of subsidence during the late spring and mid-fall (Fig. 1e and f). The extent of the deformation is similar in each epoch, with the greatest fluctuations near the city of Santa Ana and at the northwestern end of the basin. A profile shows not only that uplift mirrors the subsidence, but also reveals a net annual subsidence of about 16 mm between April 1998 and May 1999 (Fig. 1d).

Sites of likely long-term anthropogenic deformation in metropolitan Los Angeles are revealed by two-year and five-year interferograms (Fig. 2), with subsidence rates up to 34 mm yr⁻¹ and uplift rates as high as 9 mm yr⁻¹. Long-term subsidence is seen in the Santa Ana basin (6–13 mm yr⁻¹), the Wilmington (W, 28 mm yr⁻¹) and the Salt Lake (SL, 11 mm yr⁻¹) oil fields. Groundwater pumping near Chino (C)⁸ results in a subsidence rate of ~34 mm yr⁻¹. The Santa Fe Springs (SF) and portions of the Baldwin Hills (BH) oil fields are uplifting at 5–9 mm yr⁻¹. Since 1993, fluid injection has exceeded withdrawal in the Baldwin Hills oilfield, consistent with the observed uplift. The uplift mechanism for the Santa Fe Springs oilfield is unclear: the 1984–1999 extraction rates are 9% larger than the injection rates; there is no seasonal component to the uplift; and the uplift rate is much larger than the expected tectonic slip rate⁹. The Newport–Inglewood fault forms a sharp boundary to the deformation (Figs 1 and 2). The range change across the fault seasonally flips sign, indicating that the deformation is associated with groundwater pumping rather than slip along the Newport–Inglewood fault. The groundwater table is presumably truncated against this active fault.

Seasonal water table changes in the Santa Ana basin indicate an association between the aquifer management and surface deformation measured by InSAR^{10–12} and GPS (Fig. 3). Wells located near the largest seasonal range change exhibit the greatest seasonal groundwater level fluctuations, whereas wells located elsewhere in the basin have proportionally smaller groundwater elevation and range changes (Fig. 3a). The aquifer is recharged all year round and is pumped during May–September to meet the demand for water in the summer; ~78% of the water pumped from the basin in 1996–1997 came from artificial recharge¹³. On the basis of the geometry of the regional aquifer system¹⁴ and the correlation between the measured surface deformation, groundwater changes and basin

Table 1 Fault model fit to SCIGN GPS data

Faults in model	Nusfit/noise ratio*		Residual contraction†	
	All GPS sites (Fig. 5)	Reliable GPS sites (Fig. 5a)	Azimuth (° east of north)	Rate (mm yr ⁻¹)
SAF + SJF [§]	5.8	6.4	11 ± 4‡	5.8 ± 1.0‡
This study SAF + SJF	5.3	5.6	20 ± 5	5.1 ± 1.0
All strike-slip faults except PV	3.9	3.9	29 ± 5	4.2 ± 1.0
All strike-slip faults	2.3	1.6	36 ± 5	4.4 ± 0.8

* $M/N = (n^{-1} \sum (O - C)^2 / \sigma^2)^{1/2}$, where M is misfit, N is noise, n is the number of observed GPS velocities O , C is the calculated velocity, and σ is the observed error.

† Contraction azimuth was calculated from the reliable sites in Fig. 5a. Contraction rate is between sites on the Palos Verdes peninsula and sites north of downtown Los Angeles.

‡ Fault model used by ref. 6. They reported a 'north-south' contraction of 5.8 ± 1.9 mm yr⁻¹ between the Palos Verdes (PV in Fig. 5a) and San Gabriel Mountains.

The 'all strike-slip fault' model included slip along the San Andreas (SAF), San Jacinto (SJF), Elsinore, Whittier, Newport-Inglewood and Palos Verdes (PV) faults. See Fig. 5 for model description. Uncertainties; 95% confidence reported.

GPS, global positioning system.

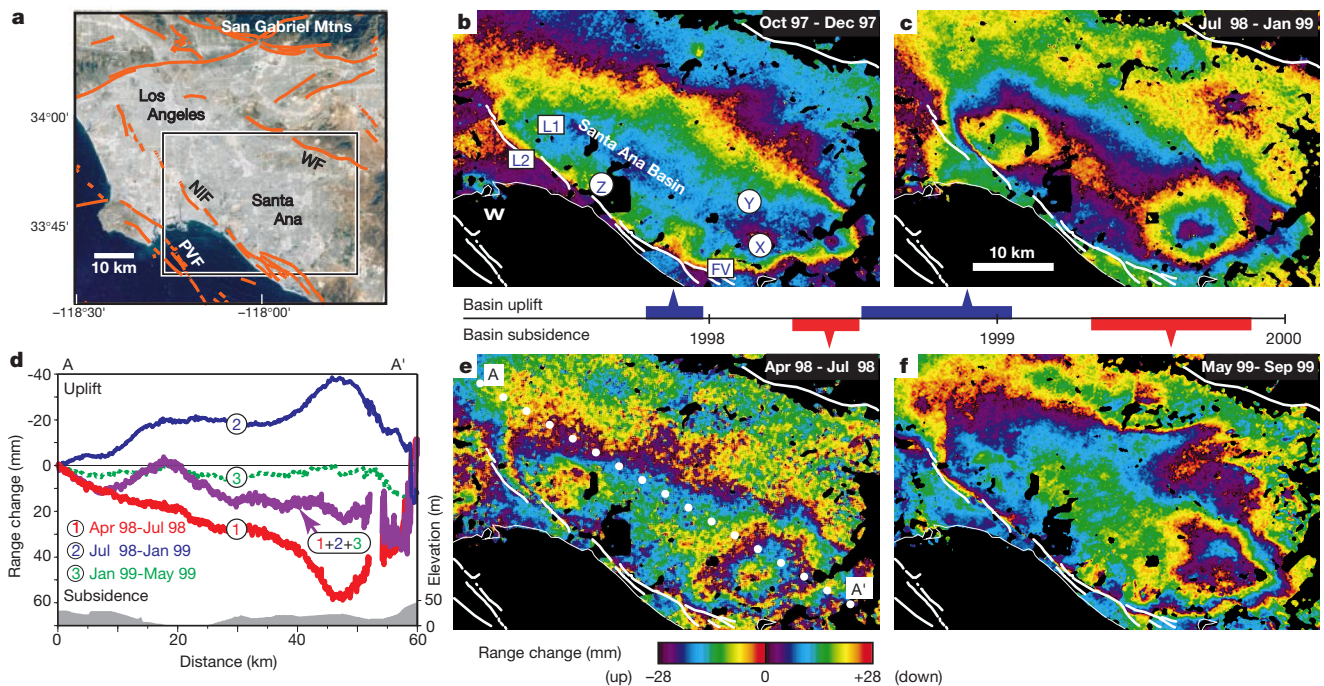


Figure 1 Seasonal deformation in the Santa Ana basin. **a**, Location map for interferograms in **b**, **c**, **e** and **f** (box), and for Figs 2 and 5 (full frame). Faults: PVF, Palos Verdes; NIF, Newport–Inglewood; and WF, Whittier. Differential interferograms (**b**, **c**, **e** and **f**) were constructed by taking the phase difference of paired synthetic aperture radar (SAR) images¹⁷, correcting for the topographic contribution using a United States Geological Survey digital elevation model and for orbital errors by removing a best-fit plane. The interferograms represent line-of-sight range changes between the surface and satellite. One cycle in colours from red through violet represents a decrease in the range of 28 mm between the ground and the satellite. The time history bar in the centre of the figure shows the period that each image spans, and summarizes the type of motion observed in the Santa Ana basin. Blue bar denotes uplift (winter months) and red bar

denotes subsidence (summer). **b**, October 1997 to December 1997. This 70-day image shows up to 34 mm of uplift in the Santa Ana basin. W, Wilmington oil field; L1, L2 and FV are GPS sites; and X, Y and Z locate water wells. **c**, July 1998 to January 1999 (175 days). The regions of maximum uplift are at the northwestern and southeastern ends of the basin, with 30 and 50 mm of uplift respectively. **e**, April 1998 to July 1998 (105 days). **f**, May to September 1999 (105 days). The maximum subsidence is 60 mm. Two regions show up to 60-mm subsidence. **d**, Unwrapped range-change profiles along the Santa Ana basin, where unwrapping corrects for phase discontinuities as the phase cycles through increments of 2. Profile location A–A' is shown as a dotted line in **e**. The deformation is independent of topography, and is thus not an artefact of elevation-dependent atmospheric delays or an inaccurate digital elevation model.

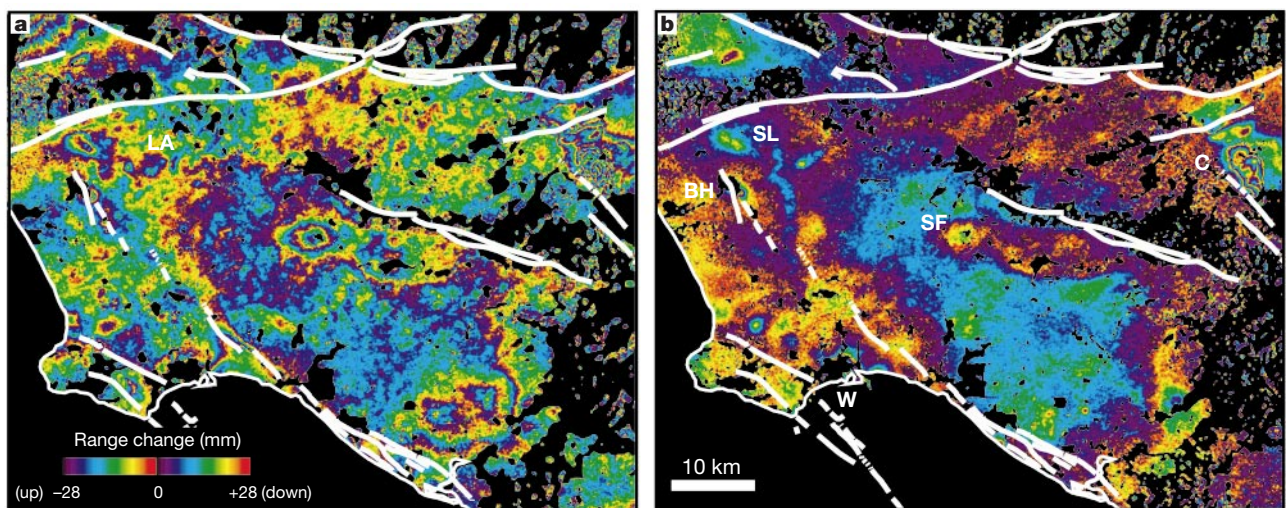


Figure 2 Long-term differential interferograms for metropolitan Los Angeles. **a**, Five-year interferogram between October 1993 and October 1998 shows over 60 mm of subsidence at the southeastern end of the Santa Ana basin during this time period (fringes in the upper left-hand corner are from the uplift associated with the 1994 Northridge earthquake). LA, downtown Los Angeles. **b**, Two-year interferogram between October 1997 and October 1998 shows ~25 mm of subsidence within the basin. BH, Baldwin

Hills; C, Chino; LA, Los Angeles; SF, Santa Fe Springs; SL, Salt Lake; and W, Wilmington oil fields. Both **a** and **b** show significant deformation near these labelled sites. The incidence angle of the ERS satellites is 23° from the vertical, thus the interferograms are most sensitive to vertical motion (the unit vector components in the look direction are: north = -0.13, east = 0.37, and up = 0.92). All data are from descending European Space Agency ERS-1 and ERS-2 satellites: track 170, frame 2925.

pumping (Fig. 3), we conclude that the seasonal deformation of the Santa Ana basin results from repeated cycles of groundwater extraction and replenishment. The long-term deformation presumably occurs from unrecoverable inelastic compaction and slow fluid depletion of the less permeable fine-grained sediments within the aquifer, in which the water level drops further after each pumping cycle.

GPS sites on the margin of the Santa Ana basin undergo seasonal horizontal motion toward and away from the basin, while sites within the basin undergo seasonal uplift and subsidence, consistent with a simple elastic model of basin behaviour. The GPS site FVPK, for example, is located on the margin of the basin and is pulled to the northeast (Fig. 3b) when the water table drops (Fig. 3a) and is pushed to the southwest when the water table rises in the winter, producing 14 mm of horizontal and 15 mm of vertical movement. Conversely, the GPS site LBC1 is located within the basin and exhibits little seasonal horizontal movement (± 3 mm) but large seasonal vertical motions (± 30 mm) (Fig. 3c). Such a pattern, with

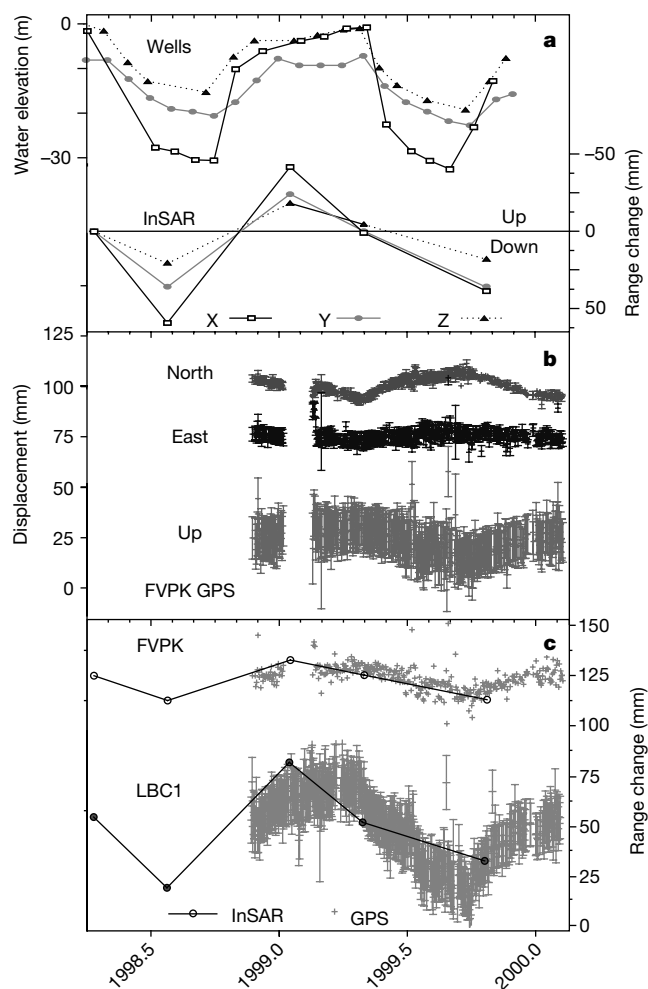


Figure 3 Comparison of groundwater fluctuations, InSAR range change, and detrended GPS time series. **a**, Seasonal elevation changes for three wells located within the Santa Ana basin (top) (located in Fig. 1b) compared with unwrapped InSAR range change time series corresponding to each well location (bottom). For example, the groundwater elevation for Well X drops ~ 30 m between the winter and summer, corresponding to subsidence of ~ 100 mm. **b**, The north, east and vertical GPS components for the site FVPK (FV in Fig. 1b). **c**, The three component GPS vector projected onto the InSAR line-of-sight vector at FVPK and LBC1 (L1 in Fig. 1b). An average linear trend has been removed from each horizontal component of the GPS time series. Water table elevations are from the Orange County water district. Well names: X, IRWD-6/1; Y, SA-35/1; and Z, SCWC-LAYT/1.

a high horizontal strain rate gradient near the periphery of an aquifer, also observed elsewhere¹⁵, can be approximated by the response of an elastic material to aquifer pumping or recharge (Fig. 4). Modelled horizontal displacements are largest near the basin margins, where they reach one-third of the subsidence over the aquifer, similar to the ratio observed in the GPS time series. Although the irregular shapes associated with the artificial subsidence and uplift features add local complexities to the horizontal deformation gradient, this elastic model can nevertheless be scaled and used as a template to help understand the observed motions in Figs 1 and 2.

About half of the continuous GPS sites in the Los Angeles basin exhibit superposed effects of tectonic motion and deformation associated with fluid pumping. After removing deformation associated with right-lateral strike-slip faults from the SCIGN velocity field, sites outside the pumping zones show a $N 36 \pm 5^\circ E$ contraction across the Los Angeles basin (Fig. 5a, Table 1). GPS sites north of Los Angeles exhibit an average southward velocity of 4.4 ± 0.8 mm yr⁻¹ relative to the Palos Verdes peninsula (PV in Fig. 5a), with no discernible contraction between the Palos Verdes and Catalina Island, 31 km to the south. In contrast, sites that are within or on the periphery of the uplift or subsidence regions have large residual velocities that are generally oriented perpendicular to the local uplift/subsidence gradient (Fig. 5b). Ironically, most of the contaminated sites lie above the Puente Hills, Elysian Park, and Compton blind thrust faults, which limits our ability to infer their geometry (Fig. 5b). Knowing which GPS sites are contaminated by artificial deformation permits either the elimination of these suspect sites or the eventual correction of the artefacts; future SCIGN sites should be placed outside artificial uplift and subsidence regions. Irrespective of the placement of GPS stations, InSAR will

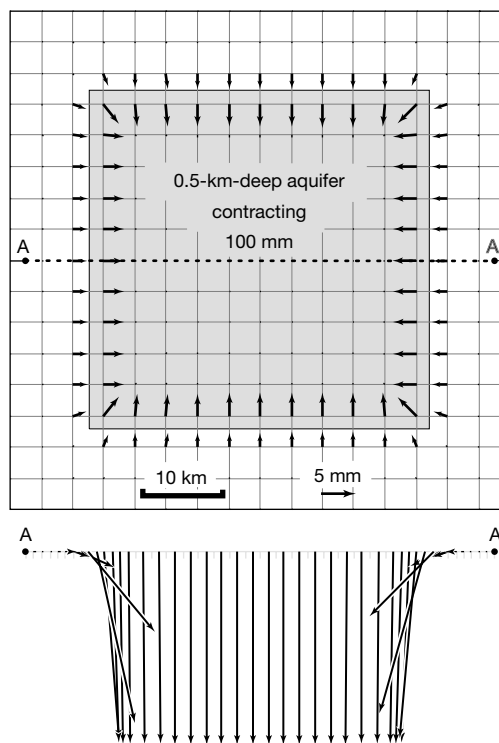


Figure 4 Simulation of basin subsidence. A thin 40 km $\times$ 40 km-aquifer is embedded at a depth of 0.5 km in an elastic halfspace with Young's modulus 10^4 MPa and Poisson's ratio 0.25, roughly approximating the Santa Ana basin. Plotted displacements (top, map view; bottom, cross-section A-A') are for uniform aquifer contraction of 100 mm owing to groundwater extraction; surface displacements would have the opposite sign for aquifer expansion owing to natural recharge or reinjection.

be needed to discriminate between tectonic and non-tectonic signals in the spatially aliased SCIGN array.

Residual GPS velocities that are free of seasonal and anthropogenic effects exhibit nearly uniaxial contraction oriented perpendicular to both the San Andreas and major strike-slip faults. The contraction rate is smaller and rotated 36° clockwise from that determined in ref. 6. This difference arises because, in addition to the San Andreas and San Jacinto faults, we also included inter-seismic slip on the Palos Verdes, Newport–Inglewood, and Whittier

faults to minimize right-lateral GPS displacement gradients (Table 1). Argus *et al.*⁶ used a smaller GPS data set to infer that regional contraction across downtown Los Angeles is confined to a 30-km-wide zone between the San Gabriel Mountains and downtown Los Angeles, but we find that the GPS sites needed to resolve the southern boundaries of the contraction region are contaminated (Fig. 5b). No more than half the geodetically observed contraction, or $<2.5\text{--}3.5\text{ mm yr}^{-1}$, may be due to thrust faulting¹⁶, with the remainder accommodated by conjugated strike-slip faults in the northern Los Angeles basin. We find not only that the residual contraction is larger than inferred¹⁶, but because the northeast oriented contractional strain rate ($56 \pm 7\text{ nanostrain yr}^{-1}$) is eight times larger than the northwest-oriented extension rate and because the contraction is oriented normal to the strike-slip faults ($N36 \pm 5^\circ E$), the residual contraction cannot be explained by unmodelled strike-slip faulting (Fig. 5a). Instead, the contraction is optimally oriented to accommodate slip on the regional blind and surface-cutting thrust faults. Further densification of the SCIGN net outside the subsidence zones may enable discrimination of competing thrust fault models, which will be vital to seismic hazard assessment in this major urban corridor. □

Received 30 January; accepted 9 July 2001.

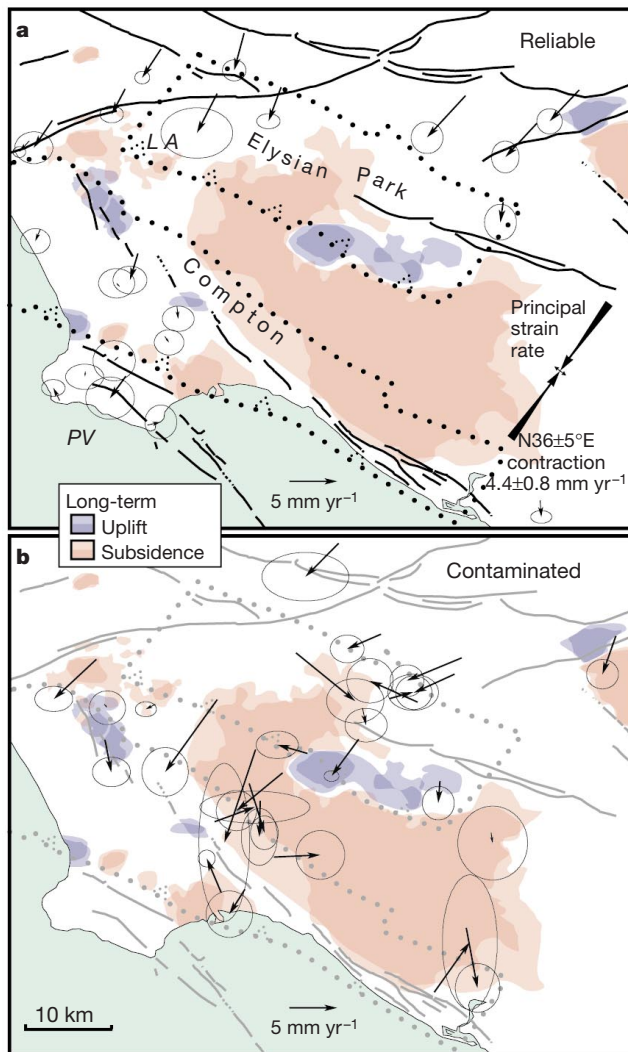


Figure 5 GPS residual velocities after removal of motion associated with strike-slip faults, superimposed on long-term uplift and subsidence features deduced from InSAR. Formal velocity uncertainties were multiplied by 4.5 to account for flicker noise in GPS time series¹⁸. The displacement field due to the San Andreas, San Jacinto and Elsinore faults was removed from the SCIGN site velocities by using a dislocation in a elastic half-space locked above 20 km depth and slipping below at the estimated Holocene slip rates¹⁹ except for the Mojave segment of the San Andreas fault, where we use 25 mm yr^{-1} (ref. 6). A 10-km locking depth was used for the Palos Verdes (2.5 mm yr^{-1}), Newport–Inglewood (1.0 mm yr^{-1}), and Whittier (2.5 mm yr^{-1}) faults. Deeper locking depths produced large misfit/noise ratios for the Holocene slip rates¹⁹. The dotted lines are the surface projections of the Elysian Park and Compton thrust faults (CT)^{9,20}. For simplicity, we do not show the segmented Puente Hill thrust fault⁹, which lies above the Elysian Park thrust with a more easterly strike. Darker shades are common to both interferograms in Fig. 2. **a**, SCIGN sites outside the margins of the aquifers and oil and gas reservoirs are classified as reliable. **b**, Sites with seasonal variations in their horizontal ($\geq 3\text{ mm yr}^{-1}$) and vertical ($\geq 8\text{ mm yr}^{-1}$) time series, or sites located near or within regions of long-term uplift or subsidence.

- Hauksson, E. & Jones, L. M. The 1987 Whittier Narrows earthquake sequence in Los Angeles, Southern California; seismological and tectonic analysis. *J. Geophys. Res.* **94**, 9569–9589 (1989).
- Hauksson, E., Jones, L. M. & Hutton, K. The 1994 Northridge earthquake sequence in California; seismological and tectonic aspects. *J. Geophys. Res.* **100**, 12335–12355 (1995).
- Dolan, J. F. *et al.* Prospects for larger or more frequent earthquakes in the Los Angeles metropolitan region. *Science* **267**, 199–205 (1995).
- Prescott, W. H. Satellites and earthquakes: A new continuous GPS array for Los Angeles, yes, it will radically improve seismic risk assessment for Los Angeles. *Eos* **77**, 417 (1996).
- Bock, Y. *et al.* Southern California Integrated GPS Network (SCIGN)/Permanent GPS Geodetic Array (PGGA). The 1996 SCEC annual report. Report No. 14-08-001-A0899 Vol. 2 (Southern California Earthquake Center, Los Angeles, 1997).
- Argus, D. F. *et al.* Shortening and thickening of metropolitan Los Angeles measured and inferred by using geodesy. *Geology* **27**, 703–706 (1999).
- Clarke, D. D., Henderson, C. P. & Strehle, R. Road log to the Long Beach area: Geologic field guide to the Long Beach area. *Ass. Petrol. Geol.* **58**, 135–159 (1987).
- Ferretti, A., Prati, C. & Rocas, F. Nonlinear subsidence rate estimation using permanent scatterers in differential SAR interferometry. *IEEE Trans. Geosci. Remote Sens.* **38**, 2202–2212 (2000).
- Shaw, J. H. & Shearer, P. M. An elusive blind-thrust fault beneath metropolitan Los Angeles. *Science* **283**, 1516–1518 (1999).
- Ireland, R. L., Poland, J. F. & Riley, F. S. Land subsidence in the San Joaquin Valley, California, as of 1980. US Geol. Surv. Open-File Rep. 82-370. 134 (US Geol. Survey, Reston, Virginia, 1982).
- Galloway, D. L. *et al.* Detection of aquifer system compaction and land subsidence using interferometric synthetic aperture radar, Antelope Valley, Mojave Desert, California. *Water Resour.* **34**, 2573–2585 (1998).
- Galloway, D., Jones, D. R. & Ingebritsen, S. E. Land subsidence in the United States. US Geol. Surv. Circ. 1182. 177 (US Geol. Survey, Reston, Virginia, 1999).
- Mills, W. *et al.* Orange County Water District, Master Plan Report. 371 (Orange County Water District, Fountain Valley, California, 1999).
- Poland, J. F. & Piper, A. M. Ground-water geology of the coastal zone, Long Beach–Santa Ana area, California. US Geol. Surv. Water-Supply Paper W1109. 162 (US Geol. Survey, Reston, Virginia, 1956).
- Carpenter, M. C. Earth-fissure movements associated with fluctuations in ground-water levels near the Picacho Mountains, south-central Arizona, 1980–84 US Geol. Surv. Prof. Paper 497-H. 58 (US Geol. Survey, Reston, Virginia, 1993).
- Walls, C. *et al.* Escape tectonics in the Los Angeles metropolitan region and implications for seismic risk. *Nature* **394**, 356–360 (1998).
- Massonnet, D. & Feigl, K. L. Radar interferometry and its application to changes in the Earth's surface. *Rev. Geophys.* **36**, 441–500 (1998).
- Zhang, J. *et al.* Southern California Permanent GPS Geodetic Array; error analysis of daily position estimates and site velocities. *J. Geophys. Res.* **102**, 18035–18055 (1997).
- Jackson, D. D. *et al.* Seismic hazards in Southern California; probable earthquakes, 1994 to 2024. *Bull. Seismol. Soc. Am.* **85**, 379–439 (1995).
- Shaw, J. H. & Suppe, J. Earthquake hazards of active blind-thrust faults under the central Los Angeles Basin, California. *J. Geophys. Res.* **101**, 8623–8642 (1996).

Acknowledgements

This work was supported by NASA Solid Earth and Natural Hazards Program. Water well data was provided by the Orange County Water District. We acknowledge the Southern California Integrated GPS Network and its sponsors, the W. M. Keck Foundation, NASA, NSF, USGS and SCEC, for providing GPS data used in this study. GPS velocity data were provided by SOPAC with assistance from Y. Bock. We thank C. Wicks for his InSAR expertise, R. O. Castle, P. Hsieh, J. Langbein, W. Prescott, F. Riley, J. C. Savage, R. Simpson and J. Shaw for reviews, and J. Shaw and A. Plesch for Puente Hills thrust fault data.

Correspondence and requests for materials should be addressed to G.W.B. (e-mail: gbawden@usgs.gov).

Evidence of intra-specific competition for food in a pelagic seabird

S. Lewis^{*†}, T. N. Sherratt^{*}, K. C. Hamer^{*} & S. Wanless[†]

^{*} Department of Biological Sciences, University of Durham, South Road, Durham DH1 3LE, United Kingdom

[†] Centre for Ecology and Hydrology, Banchory Research Station, Hill of Brathens, Aberdeenshire AB31 4BW, United Kingdom

The factors affecting the population dynamics of seabirds have long intrigued biologists^{1–5}. Current data suggest that density-dependent depletion of prey during the breeding season may regulate population size^{6–9}. However, much of the evidence for this has been circumstantial, and the underlying mechanisms are unclear^{5,10}. Here, we show that the per capita population growth rates of northern gannet *Morus bassanus* at colonies in Britain and Ireland have declined with increasing population size. Furthermore, direct observations reveal that the mean foraging trip duration of breeding gannets is positively correlated with colony size, both among colonies of different sizes in the same year, and within colonies as they change in size. To understand this phenomenon, we have developed a model which demonstrates that disturbance of fish alone can readily generate conditions under which gannets at larger colonies have to travel further to obtain food.

Some researchers have proposed that density-independent factors may keep certain seabird populations below the levels at which density-dependent factors act^{5,11}, but others have argued that populations are likely to be regulated either through density-dependent breeding success^{1,6,9} or density-dependent mortality outside the breeding season². One mechanism that may generate density-dependent population growth was identified by Ashmole¹: adults foraging close to the breeding colony are likely to cause local prey depletion, so that birds from larger colonies will have to travel further to find food for their chicks than birds from smaller colonies. In support, consistent negative correlations between colony size and numbers at neighbouring colonies have been found⁶, indicating that even seabirds foraging in areas with high marine productivity may compete for resources. However, the causes of this higher-order relationship are unclear. For instance, some authors have questioned whether seabirds could take sufficient fish to cause significant depletion¹², while others have suggested that negative correlations of this nature can at least in part be explained by a tendency of large islands to be far apart¹³.

To evaluate the validity of Ashmole's population regulation hypothesis it is essential to first establish whether seabird populations actually exhibit density-dependent growth. Although smaller black-legged kittiwake *Rissa tridactyla* colonies tend to grow proportionately faster than larger ones¹⁴, the evidence for density-dependent growth in seabird species is extremely limited. The northern gannet *M. bassanus* is a good model species both to test for density-dependent growth, and to evaluate the validity of Ashmole's specific mechanism for several reasons. First, many colonies in Britain and Ireland have been expanding in size over the past century (in part owing to a reduction of persecution¹⁵, but perhaps also owing to increased food availability¹⁶), thereby providing an ideal opportunity to test whether per capita population changes are indeed density-dependent. Second, satellite tracking of breeding birds has shown that foraging range is very closely correlated with trip duration in this species, so that the distances travelled by adults can be accurately estimated from observed trip durations^{17,18}. Finally, the gannet feeds mainly on pelagic shoaling fish^{17,19} by plunge diving for prey^{19,20}, and currently occurs in

colonies which range in size from hundreds to tens of thousands²¹, providing the opportunity to investigate directly the effects of colony size on trip duration, and hence foraging range.

To test for density-dependent population growth, historical data on the estimated population sizes of gannets at 17 of the 18 British and Irish colonies were collated (see Methods). Time series analysis²² of the longest unbroken series of yearly censuses at seven different gannetries (see Methods) identified significant density-dependent population growth at three of these colonies (Fig. 1; Bempton $b = 0.9203$, $n = 16$, $P < 0.05$; Fair Isle $b = 0.8365$, $n = 25$, $P < 0.02$; Troup Head $b = 0.7708$, $n = 11$, $P < 0.005$). More gannet colonies were censused in Britain and Ireland in 1984/85 and 1994/95 than over any other period, so we have also estimated the per capita growth rates of gannet colonies over this decade. This between-colony comparison once again showed that the per capita growth rates of colonies declined significantly with colony size (Fig. 2, $r = -0.853$, degrees of freedom, d.f. = 13, $P < 0.001$). These findings are consistent with previous observations by Nelson¹⁹, a preliminary analysis of Scottish colonies²¹ and an earlier analysis of the growth of 34 gannet colonies (R. Moss, S.W. & M. P. Harris, unpublished work).

Crucially, a comparison of observations made of chick-rearing adults in the summer of 2000 at nine gannet colonies (see Methods) showed that there was a highly significant positive correlation between trip duration (hence foraging range) and square root colony size ($r = 0.898$, d.f. = 7, $P < 0.005$, Fig. 3), with trip duration increasing threefold between the smallest and largest

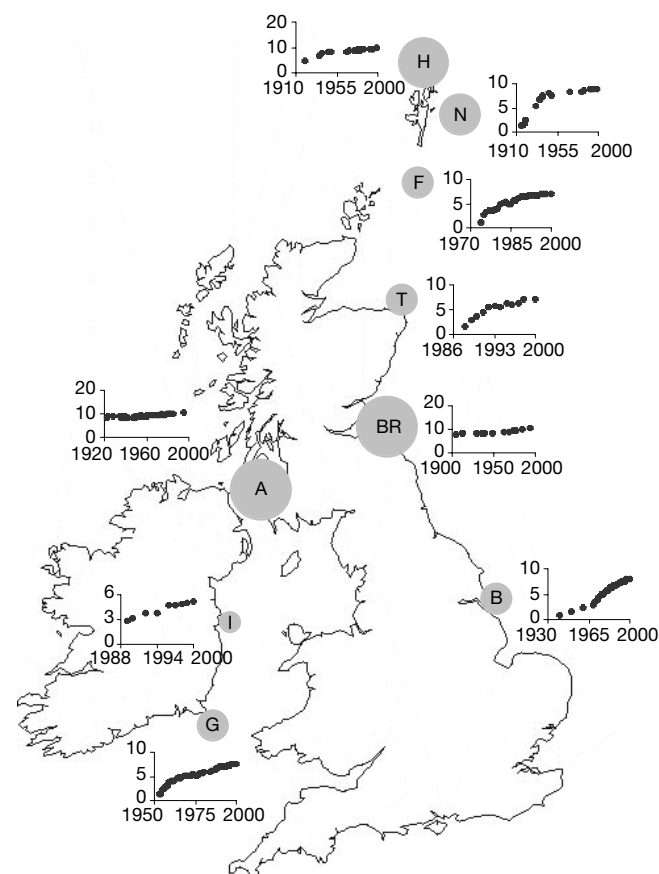


Figure 1 Location and size of gannet colonies. The current sizes, distribution and changes in gannet colony sizes over the past century at 9 of our 17 study colonies. (A, Ailsa Craig; BR, Bass Rock; B, Bempton; F, Fair Isle; G, Great Saltee; H, Hermaness; I, Ireland's Eye; N, Noss; T, Troup Head). Colony size was estimated as $\log_e$ (apparently occupied sites). The area of the circle is proportional to current colony size (see Fig. 3 for actual values). Data on trip durations were collected at these 9 colonies.

population. Although pairs at the smaller colonies spent significantly more time together at the nest (square-root-transformed data $r = 0.725$, d.f. = 7, $P < 0.05$), the rate of provisioning of chicks by the parents still tended to decrease with increasing population size (square-root-transformed data $r = 0.662$, d.f. = 7, $P = 0.052$; log-transformed data $r = -0.671$, d.f. = 7, $P < 0.05$). Historical data on foraging trip duration at four of these colonies (Fig. 3) also fitted the same regression model extremely well (2000-only regression slope is 0.0492, standard error of the mean, s.e.m. 0.00907; pooled within colony regression slope for the four colonies is 0.0622, s.e.m. 0.0102, while a linear mixed model using combined data yielded a slope of 0.0548, s.e.m. 0.0128), indicating that the relationship within colonies between years is similar to our relationship between colonies within the same year. The generality of this relationship was further confirmed using data for St Kilda (the largest colony in the North Atlantic). In 1980, the colony held 40,000 apparently occupied sites and the average trip duration was around 21 hours (S.W., unpublished work). Finally, we note that of two pairs of colonies that were very similar in size (Great Saltee (G)

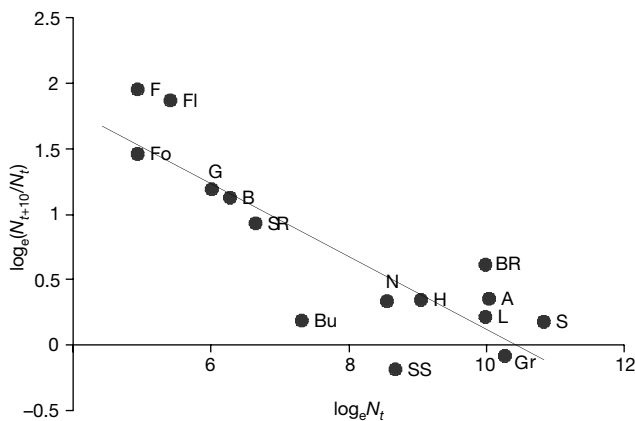


Figure 2 Density dependence in gannet colony growth. The relationship between $\log_e N_t$ (colony sizes in 1984/85) and $\log_e (N_{t+10}/N_t)$ (ratio of colony sizes in 1994/95 to 1984/85) is shown for 15 gannetries throughout the British Isles and Ireland. (Bu, Bull Rock; FI, Flannans; Fo, Foula; Gr, Grassholm; L, Little Skellig; S, St Kilda; SR, Scar Rocks; SS, Sule Stack; other colonies labelled as in Fig. 1.)

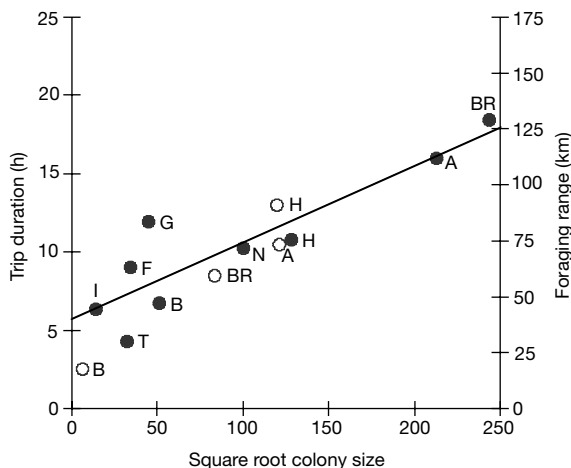


Figure 3 Gannets from larger colonies travel further to find food. The relationship between foraging trip duration and the square root of colony size for gannets is shown for the nine study colonies (labelled as in Fig. 1). Earlier data collected at four colonies (BR in 1966, B in 1972, A in 1975 and H in 1997) are represented by unfilled symbols. The line of best fit for 2000 data is shown.

and Bempton (B); Fair Isle (F) and Troup Head (T)) one was above the regression line of trip duration against colony size, and one was below it. The colonies that were above the regression line (G and F) are both approximately 50 km from other large colonies, while the colonies below the regression line (B and T) are both over 100 km away from other large colonies. This suggests that foraging birds from neighbouring colonies may to some extent compete for food.

To account for the relationship between trip duration and colony size, we have developed two complementary models in which shoals of pelagic fish show escape responses when they are attacked by diving gannets. If gannets are the primary source of disturbance and the escape response of fish involves some lateral movement, then shoals close to a colony will be attacked more frequently than those further away, and the overall tendency will be for fish shoals to diffuse outwards. Such movements readily generate a slowly growing ‘halo’¹, and we have found that under these conditions the mean foraging trip duration of gannets will increase with the square root of colony size. Yet some fish species may escape predation by swimming vertically downwards²³. Furthermore, factors such as currents or disturbance by other predatory species may effectively mix shoals to such an extent that halo patterns do not form clearly. In such cases, the geometry of central place foraging may explain our observations (Box 1). By integrating foraging rewards over a given flight path, it becomes clear that if gannets compete for available prey which is randomly distributed in discrete patches, then they should forage over approximately the same total area per bird to obtain the same amount of food, independent of colony size.

Box 1

Travel durations when central place foragers disturb prey

In this model, fish occur in discrete shoals of diameter s , which are randomly distributed with mean density d per unit distance. We assume that when shoals are disturbed by foraging gannets they become temporarily unavailable through some avoidance response (for example, sinking vertically and/or swimming away), but that they remain randomly distributed in space despite this disturbance. We consider a mainland colony of N pairs of gannets. Foraging gannets fly out to sea from the colony in straight lines with random directions, and each individual returns after they have gathered T units of fish biomass.

Let proportion p of the breeding adults leave the colony in search of food per unit time. Because gannets are in essence ‘central place’ foragers, the closer a shoal arises towards the colony the more birds are likely to encounter it. As a first approximation we have assumed that a mean proportion of $(s/\pi r)$ of gannets that leave the colony per unit time will encounter a shoal if it is at distance r from the colony (it is easy to show that the better approximation $2 \arctan(s/2r)/\pi$ yields very similar numerical results, but this function is less tractable analytically).

If g gannets encounter a given shoal (either simultaneously or in quick succession) per unit time, then let these individuals gain a mean reward of biomass $m(g)$. Here we have assumed that the immediately available prey are divided by the number of birds encountering the shoal, that is $m(g) = v/g$, where v is a constant. Nevertheless, functions with the highest foraging rewards per individual at non-zero gannet densities, such as $m(g) = 1/[g - w]^2 + c]$, where w and c are constants, also generate similar conclusions so long as peak profitability arises for relatively few gannets. Under the above conditions, an individual gannet will gain net biomass B from continuing to forage for distance j where:

$$B = \int_{r=0}^{r=j} dv/(spN/\pi r) dr$$

If gannets cease foraging when $B = T$ then they should travel distance λ where

$$\lambda = \sqrt{(2spTN/dv\pi)}$$

such that the mean total travel distance (hence travel time) will be proportional to $\sqrt{N}$.

The area covered increases with the square of the mean foraging radius, which may explain why trip duration appears to be so closely correlated with the square root of colony size.

What are the implications of adults taking longer to find food at larger colonies? One possibility is that adults have to work harder at larger colonies and therefore experience higher mortality. Perhaps more importantly, if the average food load brought back to feed the young is constant among colonies, then chicks from larger colonies will receive less food per unit time than those at smaller colonies. Many seabird species exhibit reduced breeding success during years when food is scarce^{12,24}, but gannets rear only a single chick and our analysis of available data on gannet breeding success (see Methods) did not find a significant relationship between the mean number of chicks fledged per nest and mean colony size (Spearman $r = -0.086$, d.f. = 4, $P = 0.872$). Like gannets, Brunnich's guillemot *Uria lomvia* have a single egg clutch, and while fledging success similarly did not appear to change with colony size, this species had significantly lower fledging weights in larger colonies⁷. Finally, we note that there is evidence that gannets from large colonies recruit into smaller colonies¹⁹, so it is possible that the provisioning rates within colonies could play some role in influencing where birds choose to breed for the first time.

Although significant density-dependent changes in provisioning rates are likely to be demographically important, we do not know at this stage whether feeding rates have played a major role in influencing the population growth patterns that we have observed. Other density-dependent factors may also influence gannet dynamics, such as a gradual limitation of suitable breeding sites²⁵, or perimeter-restricted growth (R. Moss, S.W. & M. P. Harris, unpublished work), and it is possible that these factors interact to some extent. Density-independent factors (such as storm events) may also affect population size¹¹, but evidently these factors were not sufficient to obscure the underlying density-dependence in this case. Lower fish densities close to a breeding colony have already been observed for one species of seabird that forages inshore⁹ but it has long puzzled biologists how seabirds might show density-dependent competition for prey at sea when seabirds appear to contribute relatively little to total fish mortality²⁶. Here we have shown that density-dependent fish disturbance is sufficient to significantly reduce the mean profitability of fish shoals close to colonies, and thereby generate suitable conditions for intra-specific competition. Because such competition can arise without significant prey mortality, our findings may have important implications for the interactions of seabirds with commercial fisheries. □

Methods

Population censuses and detection of density dependence

The estimated population sizes of gannets at 17 colonies in Britain and Ireland (A, BR, B, Bu, F, FI, Fo, G, Gr, H, I, L, N, S, SR, SS and T; see Figs 1 and 2) were derived from published counts^{19,21,27–29} of apparently occupied sites and personal data. The minimum number of censuses per colony was ten, spanning a maximum period of 1902–2000. To test for density-dependent population growth within colonies, we used the distribution-free randomization test advocated in ref. 22, which involves calculating the slope b of $\log_e N_{t+1}$ versus $\log_e \log N_t$ (where N_t is population size at time t) and comparing this with gradients derived from 1,000 randomly permuted time series. As this method relies on unbroken series of censuses at fixed intervals, we were able to use this method for the yearly censuses taken at only seven colonies (A, 1947–1985; B, 1969–1984; F, 1974–1998; G, 1964–1980; H, 1974–1980; T, 1988–1998; SR, 1968–1976). To compare per capita population growth between colonies over the same period of time, we simply calculated the proportionate increase in gannet population size at 15 colonies between 1984/85–1994/95 (I and T were not considered as they were unoccupied in 1984/85).

Estimation of trip duration

Fieldwork took place during chick rearing (overall mean age of nestlings at each colony 7.4 weeks, range 5.0–9.9 weeks) at nine of the gannet colonies in Britain and Ireland (see Fig. 3). Nestlings were always attended by at least one parent. At each colony we observed approximately 20 breedings pairs with chicks (range 18–24) for a total time of 10–55 hours. By recording the arrival and departure times of adults at these nests we estimated the overall changeover rate of gannets per nest per day. The mean trip duration at each colony (hence time between feeds) was calculated by dividing the mean time available per

day for foraging (local daylight hours minus the mean time birds are together at the nest), by the estimated changeover rate. The historical data on trip durations in Fig. 3 were collated from personal (S.W.) and published data^{19,20}. All colonies were counted in 2000 with the exception of Bass Rock, Ailsa Craig, Hermaness and Noss which were last counted in 1994, 1995, 1999 and 1999 respectively. The population sizes of these colonies for 2000 were estimated using standard population growth trajectories²¹. Foraging range was related to trip duration by the following regression equations ($r^2 = 0.94$): $D = 7.05$ (s.e.m. ± 0.22) L , where D is the maximum distance (km) and L is the trip duration (h)^{17,18}.

Determination of breeding performance

Data on the breeding success of gannets at six separate colonies between 1986 and 1999 were obtained from a Joint Nature Conservation Committee (JNCC) published report³⁰.

Received 6 June; accepted 27 June 2001.

1. Ashmole, N. P. The regulation of numbers of tropical oceanic birds. *Ibis* **103**, 458–473 (1963).
2. Lack, D. *Population Studies of Birds* 252–270 (Oxford Univ. Press, Oxford, 1966).
3. Birkhead, T. R. & Furness, R. W. in *Behavioural Ecology: Ecological Consequences of Adaptive Behaviour* (eds Sibly, R. & Smith, R.) 145–167 (Blackwell Scientific, Oxford, 1985).
4. Cairns, D. K. The regulation of seabird colony size: A hinterland model. *Am. Nat.* **134**, 141–146 (1989).
5. Croxall, J. P. & Rothery, P. in *Bird Population Studies: Relevance to Conservation and Management* (eds Perrins, C. M., Lebreton, J.-D. & Hiron, G. J. M.) 272–296 (Oxford Univ. Press, Oxford, 1991).
6. Furness, R. W. & Birkhead, T. R. Seabird colony distributions suggest competition for food supplies during the breeding season. *Nature* **311**, 655–656 (1984).
7. Gaston, A. J., Chapdelaine, G. & Noble, D. G. The growth of Thick-billed Murre chicks at colonies in Hudson Strait: inter- and intra-colony variation. *Can. J. Zool.* **61**, 2465–2475 (1983).
8. Hunt, G. L., Eppley, Z. A. & Schneider, D. C. Reproductive performance of seabirds: the importance of population and colony size. *Auk* **103**, 306–317 (1986).
9. Birt, V. L., Birt, T. P., Goulet, D., Cairns, D. K. & Montevecchi, W. A. Ashmole's halo: direct evidence for prey depletion by a seabird. *Mar. Ecol. Prog. Ser.* **40**, 205–208 (1987).
10. Wooller, R. D., Bradley, J. S. & Croxall, J. P. Long-term population studies of seabirds. *Trends Ecol. Evol.* **7**, 111–114 (1992).
11. Nur, N. & Sydeman, W. J. Survival, breeding probability and reproductive success in relation to population dynamics of Brandt's Cormorants *Phalacrocorax penicillatus*. *Bird Study* **46** (Suppl.), 575–578 (2000).
12. Croxall, J. P. *Seabirds: Feeding Ecology and Role in Marine Ecosystems* (Cambridge Univ. Press, Cambridge, 1987).
13. Forbes, L. S., Jajam, M. & Kaiser, G. W. Habitat constraints and spatial bias in seabird colony distributions. *Ecography* **23**, 575–578 (2000).
14. Coulson, J. C. The changing status of the Kittiwake *Rissa tridactyla* in the British Isles, 1969–79. *Bird Study* **30**, 9–16 (1983).
15. Lloyd, C., Tasker, M. L. & Partridge, K. *The Status of Seabirds in Britain and Ireland* 13–24 (T & AD Poyser, London, 1991).
16. Furness, R., Ensor, K. & Hudson, A. The use of fishery waste by gull populations around the British Isles. *Ardea* **80**, 105–113 (1992).
17. Hamer, K. C., Phillips, R. A., Wanless, S., Harris, M. P. & Wood, A. G. Foraging ranges, diets and feeding locations of gannets in the North Sea: evidence from satellite telemetry. *Mar. Ecol. Prog. Ser.* **200**, 257–264 (2000).
18. Hamer, K. C., Phillips, R. A., Hill, J. K., Wanless, S. & Wood, A. G. Contrasting foraging strategies of gannets *Morus bassanus* at two North Atlantic colonies: prey patchiness and foraging area fidelity. *Mar. Ecol. Prog. Ser.* (in the press).
19. Nelson, J. B. *The Sulidae: Gannets and Boobies* (Univ. Aberdeen, Oxford, 1978).
20. Garthe, S., Gremillet, D. & Furness, R. W. At-sea activity and foraging efficiency in chick-rearing northern gannets *Sula bassana*: a case study in Shetland. *Mar. Ecol. Prog. Ser.* **185**, 93–99 (1999).
21. Murray, S. & Wanless, S. The status of the Gannet in Scotland 1994–1995. *Scot. Birds* **19**, 10–27 (1997).
22. Pollard, E., Lakhani, K. & Rothery, P. The detection of density-dependence from a series of annual censuses. *Ecology* **68**, 2046–2055 (1987).
23. Logerwell, E. A. & Hargreaves, N. B. Seabird impacts on forage fish. *Proc. Int. Symp. on the Role of Forage Fishes in Marine Ecosystems* 191–196 (Alaska Sea Grant College Program Report 97-01, Univ. Alaska Fairbanks, Fairbanks, 1997).
24. Monaghan, P., Uttley, J. D. & Okill, D. Terns and sandeels: seabirds as indicators of change in marine fish populations. *J. Fish Biol.* **35** (Suppl. A), 339–340 (1989).
25. Potts, G. R., Coulson, J. C. & Deans, I. R. Population dynamics and breeding success of the shag, *Phalacrocorax aristotelis*, on the Farne Islands, Northumberland. *J. Anim. Ecol.* **49**, 465–484 (1980).
26. Furness, R. W. & Tasker, M. L. Seabird consumption in sand lance MSVPA models for the North Sea, and the impact of industrial fishing on seabird populations. *Proc. Int. Symp. on the Role of Forage Fishes in Marine Ecosystems* 147–169 (Alaska Sea Grant College Program Report 97-01, Univ. of Alaska Fairbanks, Fairbanks, 1997).
27. Fisher, J. & Vevers, H. G. The breeding distribution, history and population of the North Atlantic Gannet *Sula bassana*. *J. Anim. Ecol.* **12**, 173–213 (1943).
28. Cramp, S., Bourne, W. R. P. & Saunders, S. *The Seabirds of Britain and Ireland* 238 (Collins, London, 1974).
29. Murray, S. & Wanless, S. The status of the Gannet in Scotland 1984–1985. *Scot. Birds* **14**, 74–85 (1986).
30. Upton, A. J., Pickerell, G. & Heubeck, M. *Seabird Numbers and Breeding Success in Britain and Ireland, 1999* (UK Nature Conservation, No. 24, JNCC, Peterborough, 2000).

Acknowledgements

This paper could not have been written without the long tradition of counting gannets pioneered by J. H. Gurney and continued by J. Fisher, H. G. Vevers, D. Saunders and S. Murray. We thank T. Begg, A. Bull, J. Bull, F. Daunt, C. Gray, M. Harris, C. Holt,

M. Maher, L. Milne, D. de Palacio, K. Redman and C. Roger for recording changeover rates. We also thank D. Bates, P. Harvey, O. Merne and A. Walsh, the Neale family, D. Shaw, M. and L. Smyth, B. Zonfrillo, the Royal Society for the Protection of Birds and the Scottish Seabird Centre for logistical support. S. Bearhop, D. Elston and A. Stien gave statistical advice, while S. Albon, F. Daunt, M. Harris, R. Moss, J. Sherratt and K. Thompson commented on an earlier draft. This project was funded by a Natural Environment Research Council CASE Research Studentship to S.L.

Correspondence and requests for materials should be addressed to T.N.S. (e-mail: T.N.Sherratt@durham.ac.uk) or S.W. (e-mail: swan1@ceh.ac.uk).

Calcitic microlenses as part of the photoreceptor system in brittlestars

Joanna Aizenberg*, Alexei Tkachenko*, Steve Weiner†, Lia Addadi† & Gordon Hendler‡

* Bell Laboratories/Lucent Technologies, Murray Hill, New Jersey 07974, USA

† Department of Structural Biology, The Weizmann Institute of Science, Rehovot 76100, Israel

‡ Natural History Museum of Los Angeles County, Los Angeles, California 90007, USA

Photosensitivity in most echinoderms has been attributed to 'diffuse' dermal receptors^{1–3}. Here we report that certain single calcite crystals used by brittlestars for skeletal construction^{4,5} are also a component of specialized photosensory organs, conceivably with the function of a compound eye. The analysis of arm ossicles in *Ophiocoma*⁶ showed that in light-sensitive species, the periphery of the labyrinthine calcitic skeleton extends into a regular array of spherical microstructures that have a characteristic double-lens design. These structures are absent in light-indifferent species. Photolithographic experiments in which a photoresist film was illuminated through the lens array showed selective exposure of the photoresist under the lens centres. These results provide experimental evidence that the microlenses are optical elements that guide and focus the light inside the tissue. The estimated focal distance (4–7 μm below the lenses) coincides with the location of nerve bundles—the presumed primary photoreceptors. The lens array is designed to minimize spherical aberration and birefringence and to detect light from a particular direction. The optical performance is further optimized by phototropic chromatophores that regulate the dose of illumination reaching the receptors. These structures represent an example of a multi-functional biomaterial that fulfills both mechanical and optical functions.

Echinoderms in general, and especially the brittlestars (Ophiuroidea), exhibit a wide range of responses to light intensity, from a largely light-indifferent behaviour to pronounced colour change and rapid escape behaviour⁷. Figure 1 compares the appearance and the skeletal structure of two species of *Ophiocoma*, which represent the two extreme photosensitivity types. *Ophiocoma pumila* (Fig. 1a) shows no colour change and little reaction to illumination. *Ophiocoma wendtii* is a highly photosensitive species, and it changes colour markedly⁷, from homogeneous dark brown during the day (Fig. 1b, left) to banded grey and black at night (Fig. 1b, right). Another conspicuous behavioural response to light is negative phototaxis: *O. wendtii* is able to detect shadows and quickly escape from predators into dark crevices⁷, which they are able to identify from several centimetres away⁸. The latter reaction is particularly unexpected in these animals as the behaviour is usually associated with the presence of discrete photosensory organs. No specialized eyes have, however, been documented in brittlestars and their reactions to light have been linked to diffuse dermal receptors^{1–3}.

The sensitivity to light seems to correlate with the specialized skeletal structure of the dorsal arm plates (DAPs). These ossicles protect the upper part of each joint in brittlestar arms (Fig. 1c). Skeletal elements of echinoderms are each composed of a single crystal of oriented calcite shaped into a unique, three-dimensional mesh (stereom)^{4,5,9,10}. The diameter of the typical stereom in the DAPs of *Ophiocoma* is about 10–15 μm (Fig. 1d). In *O. wendtii* as well as in other photosensitive species⁶, the outer surface of the DAP stereom bears a characteristic array of enlarged spherical structures 40–50 μm in diameter (Fig. 1d, f). In cross-section they have a remarkably regular double-lens shape (Fig. 1g). The optical axis of the constituent calcite is oriented parallel to the lens axis and perpendicular to the plate surface⁹. The mean geometry of the lenses was inferred from the measurements of lens diameter (L) and thickness (t) in 20 random lenses sectioned through the centre (Fig. 1g):

$$t = 0.89L + 2.2 \quad (1)$$

with a correlation coefficient (r^2) of 0.91. Similar lenses were also

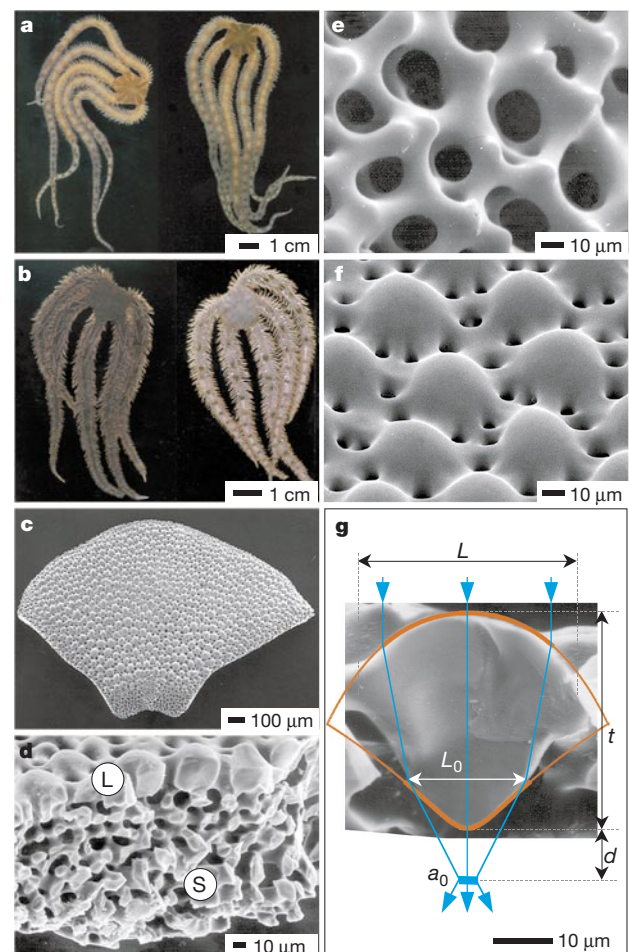


Figure 1 Appearance and skeletal structure of ophiocomid brittlestars. **a**, Light-indifferent species *Ophiocoma pumila* shows no colour change from day (left) to night (right). **b**, Light-sensitive species *O. wendtii* changes colour markedly from day (left) to night (right). **c**, Scanning electron micrograph (SEM) of a dorsal arm plate (DAP) of *O. wendtii* cleansed of organic tissue. **d**, SEM of the cross-section of a fractured DAP from *O. wendtii* showing the typical calcitic stereom (S) and the enlarged lens structures (L) that constitute the peripheral layer. **e**, SEM of the peripheral layer of a DAP of *O. pumila* showing that it lacks the enlarged lens structures. **f**, SEM of the peripheral layer of a DAP from *O. wendtii* with the enlarged lens structures. **g**, High-magnification SEM of the cross-section of an individual lens in *O. wendtii*. Red lines represent the calculated profile of a lens compensated for spherical aberration. The operational part of the calcitic lens (L_0) closely matches the profile of the compensated lens (bold red lines). The light paths are shown in blue.

found on the dorsal region of lateral arm plates. Ventral arm plates and most of the surface of lateral arm plates do not develop enlarged spherical structures, and the orientation of the optical axis of the constituent calcite is not perpendicular to the surface of the plate.

The absence of such structures on the DAPs in various relatively light-insensitive species, such as *O. pumila* (Fig. 1a, e), raises the possibility of the direct involvement of calcitic microlenses in photoreception. Calcitic microlenses were used by the trilobites^{11–13}. The presence of transparent regions of compact stereom has been reported also for sea stars¹⁴ and sea urchins¹⁵. We proposed that these calcitic microstructures might have a function in directing and focusing the light on photosensitive tissues.

To detect and visualize the lensing effect, we designed a lithographic experiment (see Fig. 2a). A DAP of *O. wendtii* was cleansed of organic tissue, and a low-magnification scanning electron micrograph (SEM) of its dorsal surface was recorded as a reference image. The inner stereom of the DAP was polished until a 43- μm -thick lens layer, free of the underlying stereom, was produced. This size was chosen to correspond to the thickness of the largest measured lens. The lens layer was placed onto a 5- μm -thick polydimethylsiloxane (PDMS) film and embedded in PDMS. PDMS served two functions: a support and a transparent organic medium with a high refractive index. A film of positive photoresist was spun on a silicon wafer. In preliminary experiments, we studied the depth of the processed photoresist as a function of exposure and set the illumination dose,

I_0 , just below the sensitivity level of photoresist (Fig. 2b, left). We performed two exposures of photoresist using the lens array embedded in PDMS as a mask^{16,17}. In the first experiment, the sample was brought directly into a conformal contact with the photoresist. In the second experiment, an additional 12- μm -thick PDMS film was used between the sample and photoresist. For our fixed exposure conditions (I_0) patterns in the photoresist are recorded only for the regions where $I > I_0$ —that is, for the areas exposed with focused light—the size of the spots in the photoresist being equal to the cross-section of the focused beam (Fig. 2b, right).

A comparative analysis of the photoresist surfaces and the reference image of the lens layer made it possible to delineate the region in the lens array (Fig. 2c) and the corresponding patterns in the photoresist obtained at the distances h_1 (48 μm) and h_2 (60 μm) from the upper surface of the lens array, respectively (Fig. 2d, e). Superimposed images of Fig. 2c–e, with the individual lenses outlined, clearly show the selective exposure of photoresist under the lens areas (Fig. 3a). This result unequivocally confirms the focusing ability of the dorsal peripheral layer of the DAP.

Figure 3b presents the rationale for the quantitative analysis of the lensing effect, by describing the focusing action of the lenses located along the dotted line in Fig. 3a. For the family of lenses that produce patterns in both photoresist layers (for example lenses 2 and 4 in Fig. 3b), we can determine distances (x) from the focal point of each

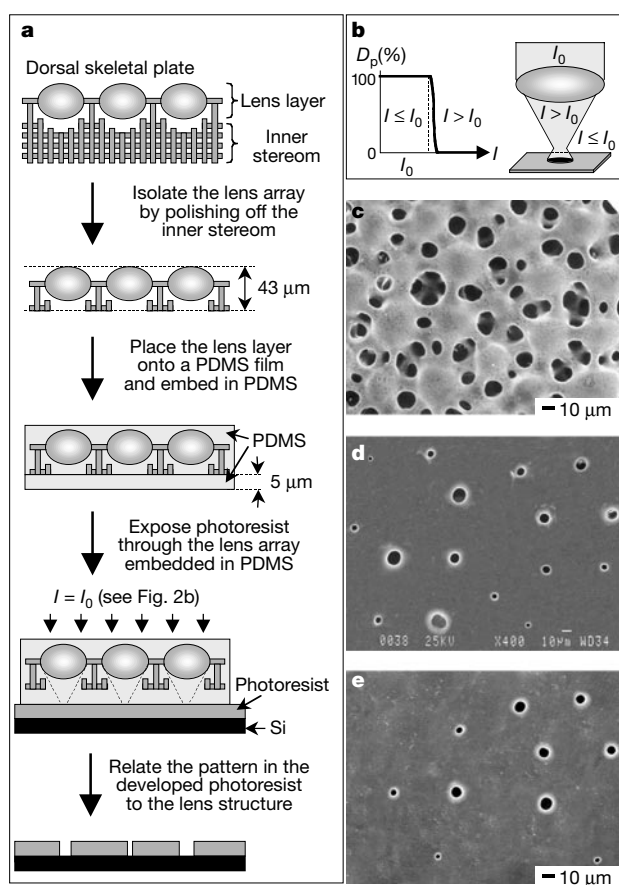


Figure 2 Lithographic experiment showing the focusing ability of the lens layer. **a**, Schematic representation of the experimental procedure (see text for details). **b**, Depth of the processed photoresist layer, D_p , as a function of exposure, I (left). The light intensity for the lithographic imaging of the focusing effect was maintained at the I_0 value, such that only areas exposed with focused light would appear in photoresist (right). **c**, SEM of the analysed area in a lens array of *O. wendtii*. **d**, **e**, SEMs of photoresist exposed at the distance h_1 (48 μm) (**d**) and h_2 (60 μm) (**e**) from the top surface of the corresponding lens array shown in **c**.

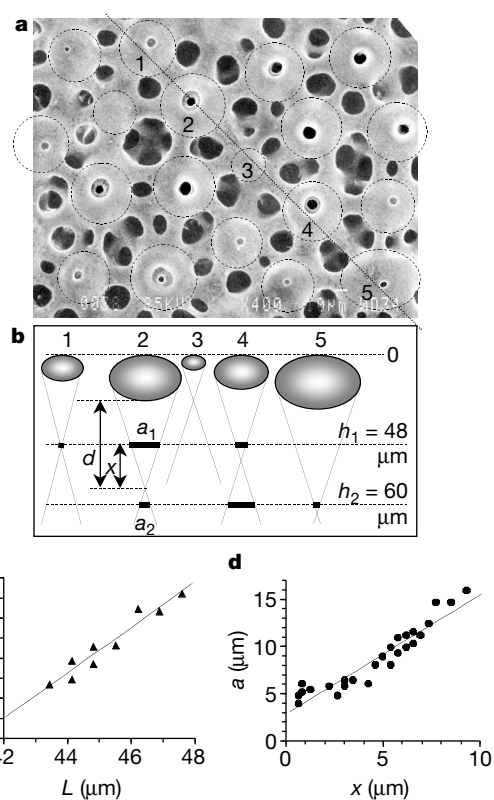


Figure 3 Analysis of the focusing effect of the lens layer. **a**, Superposition of the micrographs shown in Fig. 2c–e. Photoresist is selectively exposed under the lens areas outlined by dashed lines. **b**, Schematic representation of the focusing action of lenses located along the dotted line in **a**. Lenses with a focal plane above h_1 produce narrow wave features in the first lithographic experiment and no features in the second (lens 1). Lenses with a focal point below h_2 produce narrow features in the second experiment and no features in the first (lens 5). When the focal plane of a lens is located significantly above h_1 , no pattern in either photoresist layer is formed (lens 3). Patterns in both photoresist layers are recorded under lenses with a focal point between h_1 and h_2 (lenses 2 and 4). **c**, Distances from the focal points of lenses to the first imaging plane (x) as a function of the lens diameters (L). **d**, Sizes of the spots in photoresist (a) as a function of x .

lens to the first imaging plane:

$$x = a_1(h_2 - h_1)/(a_1 + a_2) \quad (2)$$

where a_1 and a_2 are the measured diameters of the spots recorded at the distances h_1 and h_2 from the top surface of the lens array, respectively. Figure 3c shows that the distances x depend linearly on the lens diameters L :

$$x = 1.14L - 44.8 \quad (3)$$

where r^2 is 0.93.

Equations (1) and (3) allowed us to calculate the experimental position of the focal plane for the lenses operating in PDMS:

$$d = h_1 + x - t \approx 0.25L \approx 6.5\text{--}12.5 \mu\text{m} \quad (4)$$

Using basic considerations of the geometric optics for a thick lens¹⁸ that has the characteristic shape of the microlenses, we determined the relationship between d and the effective focal length f :

$$f \approx d(1 - (1 - n_1/n_c)t/R_u)^{-1} \approx 1.5d \quad (5)$$

where R_u is the curvature radius of the upper surface of the lens, and n_1 (1.46) and n_c (1.66) are the refractive indices of PDMS and calcite along the c -axis, respectively. The position of the focal point in the actual biological environment, (d_b), can be determined using the thick-lens formula¹⁸ for two media with different refractive indices:

$$d_b \approx d(n_c/n_1 - 1)/(n_c/n_2 - 1) \approx 0.14L \approx 4\text{--}7 \mu\text{m} \quad (6)$$

where n_2 (1.34) is the refractive index of body fluids¹¹.

Figure 3d presents the sizes of the spots in the photoresist (a) as a function of x for the entire lens array. Using a linear approximation of the data, $a(x) = 1.26x + 2.9$ ($r^2 = 0.90$), one can estimate the size of the spot at the focal plane, $a_0 = a(x = 0) \approx 2.9 \mu\text{m}$, and the operational diameter of a typical lens, $L_0 = a(x = f) \approx 20 \mu\text{m}$. The intensity of the incoming light is thus enhanced at the focal point by a factor of $E = (L_0/a_0)^2 \approx 50$.

For a thick lens formed by two spherical surfaces, a_0 is related to L_0 and f by the equation¹⁸ $a_0 \approx L_0^2/f$. For such a lens, the experimental values of a_0 and f would then correspond to the maximum operational diameter $L_0 \approx 5\text{--}6 \mu\text{m}$ and the light enhancement factor $E \approx 3\text{--}4$. The experimentally determined values of L_0 and E (see above equations) are significantly higher, clearly indicating that there must exist a marked compensation for spherical aberration^{11,12,18}. To verify this conclusion, we calculated the optimal profile for the compensated calcitic lens in a biological medium (red line in Fig. 1g) and found it to be in a good agreement with the actual shape of the operational part of a lens. The proportionality of t and d to L (equations (1), (4) and (6)) suggests that all lenses are rescaled replicas of each other, implying that the entire array is compensated for aberration.

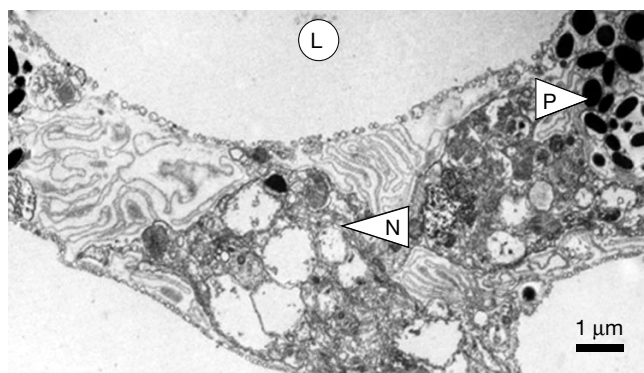


Figure 4 Transmission electron micrograph of the decalcified section of the DAP of *O. wendtii*. Sample preparation is described in ref. 6. L, lens area; N, nerve bundle; P, pigment.

If the calcitic microlenses are involved in photoreception as light guides and concentrators, one should expect the presence of receptors positioned at their focal points. Indeed, a transmission electron microscopy study⁶ of thin sections of decalcified DAPs revealed bundles of nerve fibre located at the predicted distance d_b beneath the lens layer (Fig. 4). The diameters of the neural bundles ($b \approx 2\text{--}4 \mu\text{m}$) correspond well to the estimated size of the focused spot a_0 . Furthermore, what is seen as a diurnal colour change of *O. wendtii* (Fig. 1b) is, in fact, a filtering and diaphragm action of chromatophores. These chromatophores regulate the intensity of light reaching the lenses by extending their pigment-filled processes to cover the lens during the day and retracting them to a lateral position between the lenses during the night⁶ (Fig. 4). Although the characteristics of echinoderm dermal photoreceptors and their precise locations are poorly understood and controversial^{3,6,19,20}, the likelihood of sub-lens photoreceptors in *O. wendtii* also corresponds with neurophysiological^{21,22} and biochemical²⁰ data.

On the basis of our results, we suggest that the array of calcitic microlenses with their unique focusing effect and underlying neural receptors may form a specialized photoreceptor system with a conceivable compound-eye capability. For a compound eye to operate, each unit (lens plus photoreceptor) must respond only to light coming from a single direction²³. The angular selectivity is determined by three principal parameters: first, the angular resolution of the lens, $\phi = a_0/f$. Our experiments yield a ϕ -value of $\sim 10^\circ$. The second parameter is the ratio of the detector diameter to the focal length, m . Although we do not know the effective size of the receptors in *O. wendtii*, we can estimate the maximum value of m that is limited by the diameter of the sub-lensar nerve bundles: $m_{\max} = b/f \approx \phi$. The last parameter that determines the angular selectivity is the alignment of the detector with respect to the optical axis. From our observation that the lenses are compensated for aberration, it follows that when light deviates from the optimal direction by ψ , the spot size increases linearly with $|\psi|$: $a(\psi) - a_0 \approx |\psi|L_0$. This implies that the focal intensity decays as $(|\psi| + \psi^*)^{-2}$, that is, it is sharply peaked in the vicinity of the optimal incident angle, with the characteristic width of the peak ($\psi^* \approx a_0/L_0$) being of the order of the angular resolution ϕ . This effect suggests a plausible mechanism for targeting specific receptors aligned with the lens axis, on the basis of significantly stronger receptor response to the incoming focused light.

The selected nerve bundle will then efficiently detect signals that come only from one direction with the angular selectivity of $\sim 10^\circ$, making the calcitic microlenses suitable for a compound-eye unit. It is doubtful, however, that one DAP could function as a single eye—even though it has a slightly curved shape, the individual lenses are all oriented along the crystallographic c -axis, minimizing the effect of birefringence. This suggests that the entire array is a highly redundant optical element free of aberration that detects the light from a particular direction. As a brittlestar has a large number of DAPs and additional arrays of lenses on the lateral surfaces, all of which are oriented differently and are capable of changing position as the arm moves, it could potentially extract a considerable amount of visual information about its environment. Although we have only limited evidence that the lens apparatus operates at a distance⁸, which would conform to the definition of an eye, our results suggest that the presence of such structures may be sufficient to elicit the rapid, coordinated behaviours, such as detection of predators and retreat toward crevices, that imply the occurrence of vision. These are the very abilities that are central to the survival of individuals of *O. wendtii* in their natural habitat⁷.

The demonstrated use of calcite by brittlestars, both as an optical element and as a mechanical support, illustrates the remarkable ability of organisms, through the process of evolution, to optimize one material for several functions, and provides new ideas for the fabrication of 'smart' materials^{24,25}. □

Received 29 March; accepted 25 June 2001.

- Hyman, L. H. *The Invertebrates* Vol. 4 (McGraw-Hill, New York, 1955).
- Millott, N. The photosensitivity of echinoids. *Adv. Mar. Biol.* **13**, 1–52 (1975).
- Yoshida, M., Takasu, N. & Tamotsu, S. In *Photoreception and Vision in Invertebrates*; (ed. Ali, M. A.) 743–771 (Plenum, New York, 1984).
- Lowenstam, H. A. & Weiner, S. *On Biomineralization* 123–134 (Oxford Univ. Press, Oxford, 1989).
- Wainwright, S. A., Biggs, W. D., Currey, J. D. & Gosline, J. M. *Mechanical Design in Organisms* (John Wiley, New York, 1976).
- Hendler, G. & Byrne, M. Fine structure of the dorsal arm plate of *Ophiocoma wendti* (Echinodermata, Ophiuroidea). *Zoomorphology* **107**, 261–272 (1987).
- Hendler, G. Brittlestar color-change and phototaxis (Echinodermata: Ophiuroidea: Ophiocomidae). *Mar. Ecol.* **5**, 379–401 (1984).
- Cowles, R. P. Stimuli produced by light and by contact with solid walls as factors in the behavior of ophiuroids. *J. Exp. Zool.* **9**, 387–416 (1910).
- Donnay, G. & Pawson, D. L. X-ray diffraction studies of echinoderm plates. *Science* **166**, 1147–1150 (1969).
- Amey, L., Hermann, R., Wilt, F. & Dubois, P. Ultrastructural localization of proteins involved in sea urchin biomineralization. *J. Histochem. Cytochem.* **47**, 1189–1200 (1999).
- Clarkson, E. N. K. & Levi-Setti, R. Trilobite eyes and the optics of Des Cartes and Huygens. *Nature* **254**, 663–667 (1975).
- Gal, J., Horvath, G., Clarkson, E. N. K. & Haiman, O. Image formation by bifocal lenses in a trilobite eye? *Vision Res.* **40**, 843–853 (2000).
- Towe, K. M. Trilobite eyes: calcified lenses *in vivo*. *Science* **179**, 1007–1010 (1973).
- Döderlein, L. Ueber 'Kryallkörper' bei Seesternen. *Denkschr. Med. Nat. Ges. Jena* **8**, 491–494 (1898).
- Smith, A. B. In *Special Papers in Palaeontology* **25**, 1–81 (Palaeontology Association, London, 1980).
- Xia, Y. N. & Whitesides, G. M. Soft lithography. *Annu. Rev. Mater. Sci.* **28**, 153–184 (1998).
- Aizenberg, J., Rogers, J. A., Paul, K. E. & Whitesides, G. M. Imaging the irradiance distribution in the optical near field. *Appl. Phys. Lett.* **71**, 3773–3775 (1997).
- Flint, H. T. *Geometrical Optics* (Methuen, London, 1936).
- Cobb, J. L. S. & Moore, A. Comparative studies on receptor structure in the brittlestar *Ophiura ophiura*. *J. Neurocytol.* **15**, 97–108 (1986).
- Johnsen, S. Identification and localization of a possible rhodopsin in the echinoderms *Asterias forbesi* (Asteroidea) and *Ophiocoma brevispinum* (Ophiuroidea). *Biol. Bull.* **193**, 97–105 (1997).
- Cobb, J. L. S. & Hendler, G. Neurophysiological characterization of the photoreceptor system in brittlestars. *Comp. Biochem. Physiol.* **97A**, 329–333 (1990).
- Stubbs, T. R. In *Echinoderms* (ed. Lawrence, J. M.) 403–408 (Proc. Int. Echinoderms Conf., Tampa Bay, 1982).
- Land, M. F. In *Comparative Physiology and Evolution in Invertebrates B: Invertebrate Visual Centers and Behavior I* (ed. Autrum, H.) 471–592 (Springer, Berlin, 1981).
- Mann, S. & Ozin, G. A. Synthesis of inorganic materials with complex form. *Nature* **382**, 313–318 (1996).
- Belcher, A. M., Hansma, P. K., Stucky, G. D. & Morse, D. E. First steps in harnessing the potential of biomineralization. *Acta Mater.* **46**, 733–736 (1998).

Acknowledgements

We thank P. Wiltzius and M. Megens for helpful discussions.

Correspondence and requests for materials should be addressed to J.A. (e-mail: jaizenberg@lucent.com).

Delineation of prognostic biomarkers in prostate cancer

Saravana M. Dhanasekaran*, **Terrence R. Barrette***, **Debashis Ghosh†**, **Rajal Shah***, **Sooryanarayana Varambally***, **Kotoku Kurachi‡**, **Kenneth J. Pienta§¶**, **Mark A. Rubin*§¶** & **Arul M. Chinnaiyan*§¶**

Departments of * Pathology, † Biostatistics, ‡ Human Genetics, § Urology, ¶ Internal Medicine, and ¶ Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA
These authors share senior authorship

Prostate cancer is the most frequently diagnosed cancer in American men^{1,2}. Screening for prostate-specific antigen (PSA) has led to earlier detection of prostate cancer³, but elevated serum PSA levels may be present in non-malignant conditions such as benign prostatic hyperplasia (BPH). Characterization of gene-expression profiles that molecularly distinguish prostatic neoplasms may identify genes involved in prostate carcinogenesis, elucidate clinical biomarkers, and lead to an improved classification of prostate cancer^{4–6}. Using microarrays of complementary DNA, we

examined gene-expression profiles of more than 50 normal and neoplastic prostate specimens and three common prostate-cancer cell lines. Signature expression profiles of normal adjacent prostate (NAP), BPH, localized prostate cancer, and metastatic, hormone-refractory prostate cancer were determined. Here we establish many associations between genes and prostate cancer. We assessed two of these genes—hepsin, a transmembrane serine protease, and pim-1, a serine/threonine kinase—at the protein level using tissue microarrays consisting of over 700 clinically stratified prostate-cancer specimens. Expression of hepsin and pim-1 proteins was significantly correlated with measures of clinical outcome. Thus, the integration of cDNA microarray, high-density tissue microarray, and linked clinical and pathology data is a powerful approach to molecular profiling of human cancer.

We developed a 9,984-element (10K) human cDNA microarray to analyse gene expression profiles in benign and malignant prostate tissue. As with previous cancer profiling studies^{7–10}, molecular classification of prostate cancer was one of the goals of this analysis. We used two distinct reference samples for comparative microarray analysis: NAP tissue from patients with prostate cancer, and prostate tissue from men without documented prostate pathology. By making direct comparisons against normal tissue counterparts, we took advantage of a 'subtractive' effect, which emphasized genes that consistently distinguished normal and neoplastic tissues.

Prostate tissues used in microarray analysis included 4 BPH samples, 8 NAP samples, 1 commercial pool of normal prostate tissue (from 19 individuals), 1 prostatitis sample, and 11 localized and 7 metastatic prostate-cancer samples. Three cell lines from metastatic prostate cancer (DU-145, LnCAP and PC3) were also profiled for gene expression. Twenty-eight additional prostate tissue specimens were profiled and the data included in the Supplementary Information (samples of 9 BPH, 1 NAP, 13 metastatic and 5 localized prostate cancers). Fluorescently labelled (Cy5) cDNA was prepared from total RNA from each experimental sample. A second distinguishable fluorescent dye (Cy3) was used to label the two reference samples used in this study: a pool of NAP from four independent patients with prostate cancer and a commercial pool of normal prostate tissues. A direct comparison between the NAP and commercial pools was also made and notable differences in gene expression were readily apparent (Fig. 1 and Supplementary Information).

In all, more than 80 cDNA microarrays were used to assess gene expression in four clinical states of prostate-derived tissues and two distinct reference pools of normal specimens. Figure 1 provides an overview of the variation in gene expression across the different tissue specimens analysed (the full data set and 28 further samples can be seen in the Supplementary Information). A hierarchical clustering algorithm was used to group genes and experimental samples on the basis of similarities of gene expression over all that were tested. Relationships between the experimental samples are summarized as dendrograms (Fig. 1a), in which the pattern and length of the branches reflect the relatedness of the samples. Benign conditions of the prostate, such as BPH and NAP, cluster separately from malignant prostate-cancer cell lines or tissues, regardless of the reference pool used. Within the prostate-cancer cluster, metastatic and clinically localized prostate cancer formed distinct subgroups.

Eisen matrix formats¹¹ of the variation in gene expression show clusters of coordinately expressed genes, highlighting relationships between specimens (black bars in Fig. 1b, c). For example, clusters B3 and C1 represent genes downregulated in both localized and metastatic prostate cancer (Fig. 1b, c). By contrast, clusters B6 and B4 highlight genes that are specifically up- or downregulated in metastatic prostate cancer, respectively (Fig. 1b). *IGFBP-5*, *DAN1*, *FAT* tumour suppressor and *RAB5A* are examples of genes that are downregulated specifically in metastatic prostate cancer and also have a proposed role in oncogenesis (magnified regions, Fig. 1b).

Similarly, cancer-related genes that are upregulated in metastatic prostate cancer include *MTA-1* (metastasis-associated 1), *MYBL2* and *FLS353* (preferentially expressed in colorectal cancer).

We used several methods of gene selection to create a more limited set of genes for future exploration. In one method (others can be found in the Supplementary Information), we computed *t*-statistics (of prostate cancer versus benign tissue) for each gene. First, the *t*-values were ranked by absolute magnitude, which takes into account the inter-sample variability in expression ratios. Second, they were ranked by the magnitude of the numerator of the test statistic, which is based on the biological difference in expression ratios and designated as 'effect size'. We examined a list of the genes with the 200 largest effect sizes and 200 largest *t*-values and a scatterplot of their values (see Supplementary Information Fig. 5) to identify candidate genes. Implementing this methodology on both reference-pool data sets yielded genes

that included *HPN* (*hepsin*), *PIM1*, *LIM* (*ENIGMA*), *TIMP2*, *HEVIN*, *RIG* and *THBS1* (*thrombospondin-1*), among others. Several genes are covered in detail in the Supplementary Information.

Many of the genes identified in these 'focused' clusters have been implicated directly or indirectly as cancer biomarkers or mediators of carcinogenesis. For example, the tumour-suppressor gene *PTEN* was downregulated, whereas the proto-oncogene *MYC* was upregulated in our microarray analysis of prostate cancer (see ref. 1 and Supplementary Information). Likewise, we observed decreased expression of *CDH1* (*E-cadherin*) and increased expression of fatty acid synthase, both of which have been shown to be dysregulated in prostate cancer^{12,13}. In addition to uncharacterized expressed sequence tags (ESTs), numerous genes were identified by our screen but not previously known to be associated with prostate cancer. Furthermore, we assessed the data by examining functional

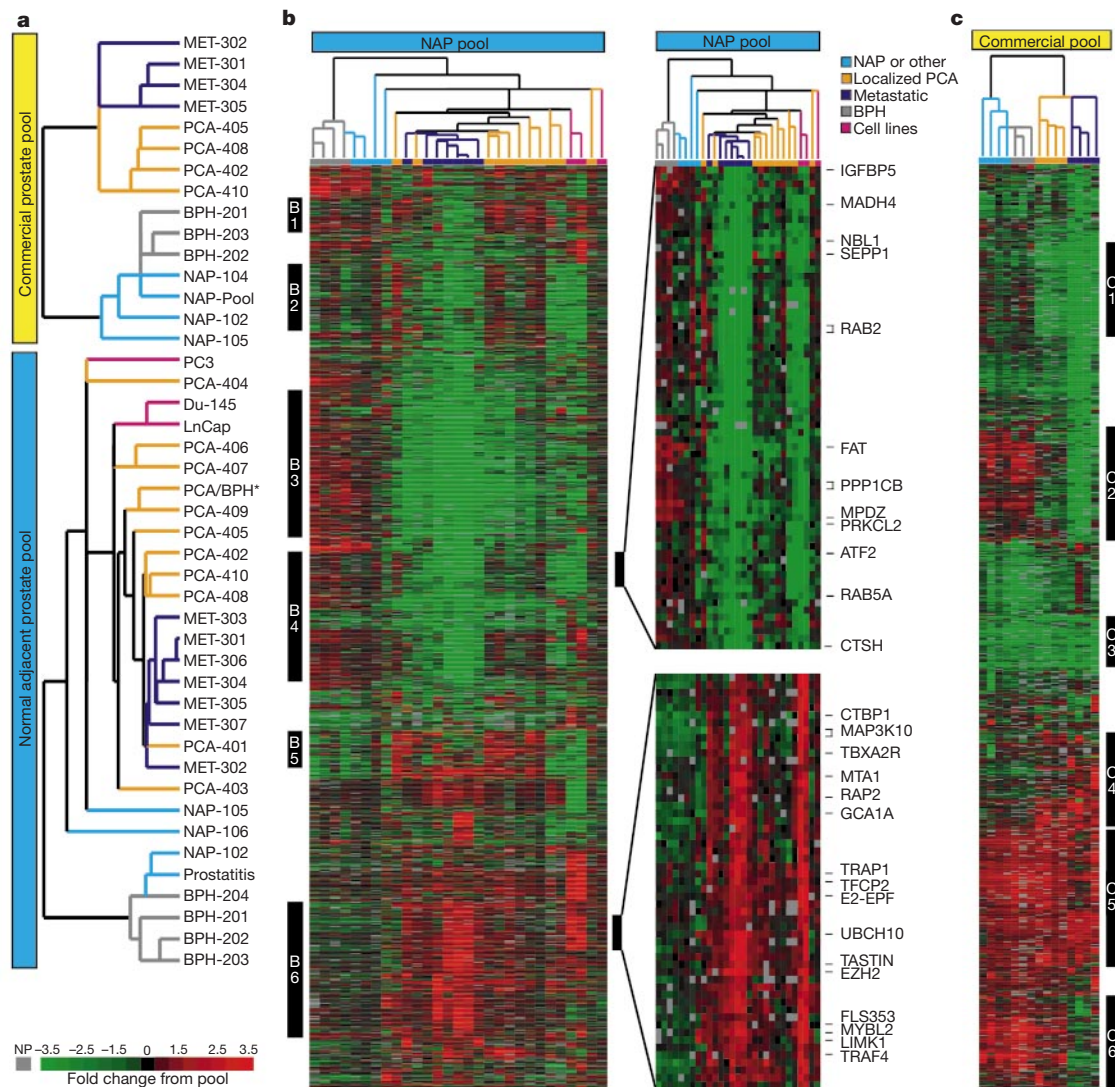


Figure 1 The molecular signature of prostate cancer. Gene-expression profiles of 44 experimental samples were performed using two distinct normal prostate reference pools (NAP pool and commercial pool). Samples analysed included NAP, localized (PCA) and metastatic (MET) prostate cancer, BPH, three prostate-cancer cell lines (DU-145, LnCAP and PC3) and prostatitis. **a**, Dendrograms reveal the variation in gene-expression pattern between experimental samples in the two pools. Asterisk indicates a sample that was initially documented as BPH, but later confirmed to have 5% cancer tissue. **b**, **c**, Cluster diagrams of the sample groups compared with the pool of normal adjacent prostate (**b**) or of commercial prostate (**c**). This clustering identifies distinct patterns of gene expression

among the groups. Each row represents a single gene (1,520 genes in **b**, and 1,006 genes in **c**). Data are the quotient of hybridization of the fluorescent cDNA probe prepared from each experimental messenger RNA to its reference pool, and are a measure of relative gene expression in each sample. Red and green represent up- and downregulation, respectively (see scale in lower left corner), relative to the median of the pool. Grey signifies technically inadequate or missing data (NP, not present). Black bars on the sides (B1–B6 and C1–C6) indicate regions with characteristic gene expression signatures. Enlarged sections of B4 and B6 show selected genes. See Supplementary Information for full data sets.

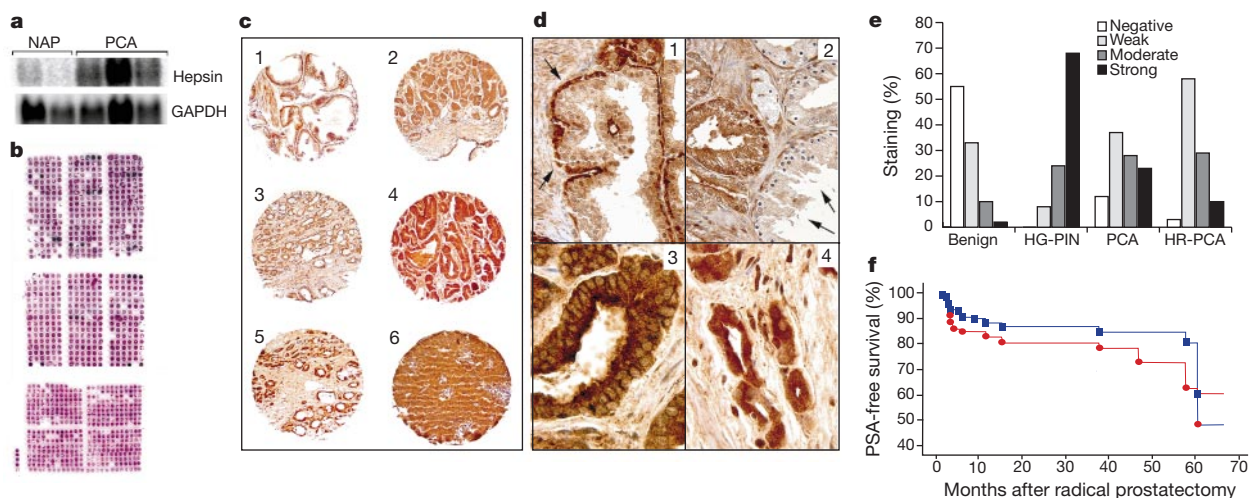


Figure 2 Hepsin is overexpressed in prostate cancer. **a**, Northern blot analysis of human hepsin and normalization with GAPDH. PCA, localized prostate cancer. **b**, Tissue microarrays used for hepsin analysis (stained with haematoxylin and eosin). **c**, Representative elements of a tissue microarray stained with anti-hepsin antibody. Immunohistochemical stains demonstrate absent or weak staining of benign prostate (c1), and strong staining in localized prostate cancer (c2–6). Magnification $\times 100$. **d**, Benign prostate glands demonstrate strong basal cell staining (d1, arrows) but weak expression in

the secretory luminal cells (d2, arrows). Where prostate cancer and benign prostate glands are seen (d2), marked differences in hepsin staining are observed. Infiltrating prostate cancers (d3 and d4) demonstrate strong expression of hepsin. Magnification $\times 400$. **e**, Histogram of hepsin expression by tissue type. The relative strength of hepsin staining was assessed qualitatively (see text). HR-PCA, hormone refractory prostate cancer. **f**, Kaplan–Meier analysis. PSA-free survival was stratified by level of hepsin expression into two categories: absent or low (circles), and moderate or strong (squares).

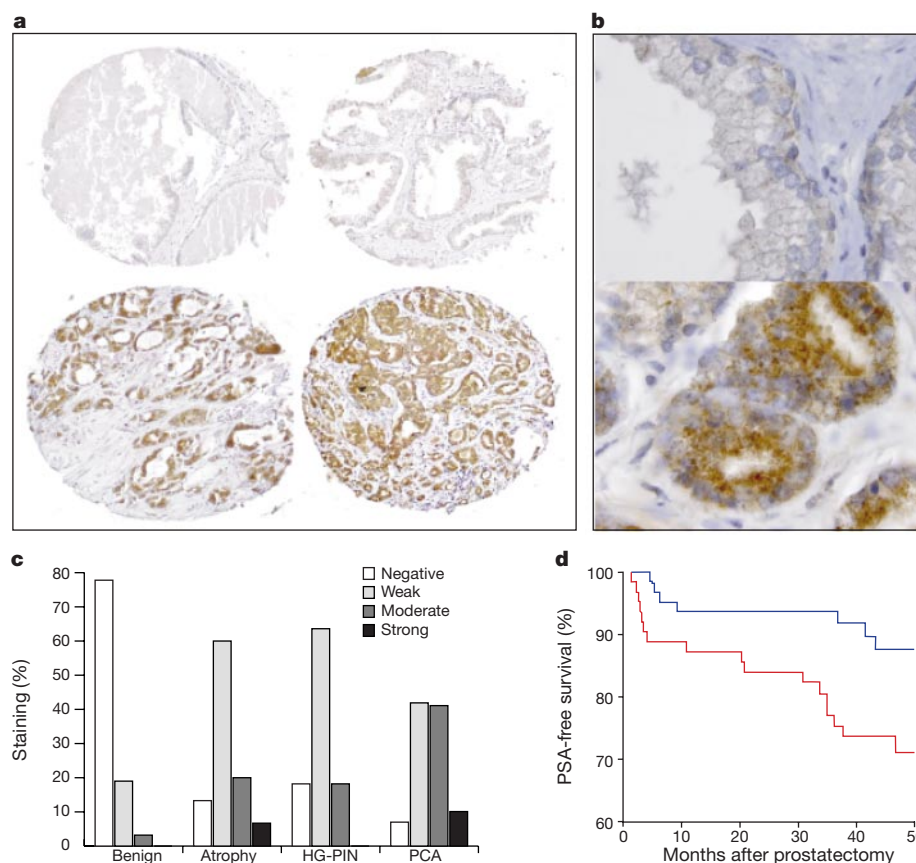


Figure 3 PIM1 is overexpressed in prostate cancer. **a**, Representative elements of a tissue microarray stained with anti-PIM1 antibody. Staining is absent or weak in benign prostate (top), but strong in the cytoplasm of localized prostate cancer (bottom). Magnification $\times 200$. **b**, PIM1 expression is absent or weak in the secretory luminal cells of benign prostate glands (top), but strong in infiltrating prostate cancers (bottom). Magnification

$\times 1,000$. **c**, PIM1 protein expression by tissue type as assessed from 810 tissue-microarray elements. **d**, Kaplan–Meier analysis shows that patients with prostate cancer that have negative or weak PIM1 expression (red) are at a greater risk of developing PSA failure after prostatectomy than those that have moderate or strong expression (blue) (log-rank test, $P = 0.04$).

groups of known, named genes (see Supplementary Information).

Selected genes identified by microarray analysis were corroborated by northern analysis. For example, hevin, four and a half LIM domain protein and gelsolin were shown to be 3.2-, 3.2- and 1.9-fold downregulated, respectively, by microarray and 8.8-, 4.5- and 3.5-fold downregulated by northern analysis (data not shown). Similarly, hepsin was 4.3-fold upregulated by microarray and 11.3-fold upregulated by northern analysis (Fig. 2a). As hepsin is a cell-surface serine protease with transcript expression precisely restricted to localized and metastatic prostate cancer, we chose to examine its expression in more detail at the protein level.

We used a method for evaluating tumour tissues in large numbers on a single glass slide¹⁴. Sections of the microarrays of prostate-cancer tissue¹⁵ used in this study are shown stained with haematoxylin and eosin in Fig. 2b. Hepsin immunohistochemistry was performed on three tissue microarrays containing a total of 738 specimens (Fig. 2c–e) using an affinity-purified hepsin-peptide antibody¹⁶. These samples included benign prostate tissue, BPH, high-grade prostatic intra-epithelial neoplasia (HG-PIN), and localized and metastatic prostate-cancer samples. Overall, hepsin exhibited staining predominantly in the plasma membrane and was preferentially expressed in neoplastic prostate over benign prostate (Mann–Whitney *U*-test, $P < 0.0001$). Importantly, the precursor lesion of prostate cancer, HG-PIN, had the strongest expression of hepsin, and almost never lacked staining (Mann–Whitney *U*-test, $P < 0.0001$). Most cases of low or absent hepsin staining were seen in benign prostate specimens (Fig. 2e). Interestingly, hormone-refractory metastatic cancers were intermediate in staining intensity between localized prostate tumours and benign prostate.

Men who develop elevated PSA levels following radical prostatectomy are at a high risk to develop distant metastases and die because of prostate cancer (termed PSA failure)¹⁷. Therefore to assess the usefulness of hepsin as a potential prostate-cancer biomarker, PSA failure was defined as a PSA elevation of greater than 0.2 ng ml^{-1} following radical prostatectomy. Outcomes analysis was performed on like-treated cases, and therefore was restricted to the 334 prostate-cancer samples from 78 men with clinically localized prostate cancer (each patient's tumour is replicated on average four times). The analysis was performed on 334 samples of localized prostate cancer, treating each as an independent sample. PSA elevation following radical prostatectomy was significantly associated with absent or low hepsin immunostaining, (Fig. 2f; log-rank test, $P = 0.03$).

Multivariate analysis was performed to examine whether these results were independent of the Gleason score, a well established histological grading system for prostate cancer¹⁸. Fitting a Cox proportional hazards model indicates an association of weak or absent hepsin protein expression in prostate cancer with increased risk of PSA elevation following prostatectomy, similar to having a high Gleason score (corresponding hazard ratios (HRs) were 2.9 ($P = 0.0004$) and 1.65 ($P = 0.037$), respectively). Weak or absent hepsin expression was also associated with large prostate cancers: the median tumour dimension for prostate tumours with moderate or strong expression was 1.3 cm but 1.5 cm for tumours with absent or weak staining (Mann–Whitney *U*-test, $P = 0.043$). Taken together, expression of hepsin protein in prostate cancer correlated inversely with measures of patient prognosis.

It is well known that the oncogene *MYC* is overexpressed in prostate cancer¹⁹; we demonstrate here that the protein kinase encoded by the proto-oncogene *PIM1* is similarly upregulated (Supplementary Information Fig. 7). Our analysis supports a remarkably similar co-transcriptional regulation (or gene amplification) of *PIM1* and *MYC*, possibly mediating a synergistic oncogenic effect in prostate cancer. Similar to hepsin, we explored expression of the protein PIM1 kinase in prostate cancer using high-density tissue microarrays. No or weak PIM1 expression was observed in most benign prostatic epithelial (97%), prostatic

atrophy (73%) and HG-PIN (82%) samples (Fig. 3a–c). In contrast, moderate to strong PIM1 expression was observed in approximately half of the prostate cancer samples (51%; Fig. 3c). Kaplan–Meier analysis for PSA-free survival demonstrated that positive extraprostatic extensions, seminal vesicle invasions, Gleason scores greater than seven and decreased PIM1 expression were associated with a higher cumulative rate of PSA failure (Fig. 3d). By univariate Cox models, PIM1 expression is a strong predictor of PSA recurrence (HR = 2.1 (95% confidence interval 1.2–3.8), $P = 0.01$). Among the variables examined, significant predictors of PSA recurrence were Gleason score (HR = 1.8 (1.1–3.0), $P = 0.03$), Gleason pattern 4/5 prostate cancer (HR = 3.9 (1.8–8.3), $P < 0.001$), extraprostatic extension status (HR = 2.6 (1.6–4.2), $P < 0.0001$), surgical margin status (HR = 2.6 (1.2–5.6), $P = 0.01$), seminal vesicle status (HR = 3.5 (2.0–6.2), $P < 0.0001$), the natural log of pre-operative PSA level (HR = 2.5 (1.6–3.8), $P < 0.001$), and maximum tumour dimension (HR = 2.7 (1.6–4.7), $P < 0.0001$). By a multivariate Cox model, significant predictors of PSA recurrence were the presence of Gleason grades 4 and 5 prostate cancer (HR = 3.8 (1.4–10.0), $P < 0.01$), the natural log of pre-operative PSA level (HR = 2.1 (1.1–3.9), $P = 0.02$), and decreased PIM1 expression (HR = 4.5 (1.6–15.2), $P = 0.01$). Thus, even more so than hepsin, decreased expression of PIM1 kinase in prostate cancer correlated significantly with measures of poor patient outcome.

Analysis of the differential gene-expression profiles of normal and neoplastic prostate has identified a select set of genes that define a molecular signature for prostate cancer. By making direct comparative hybridizations of normal and neoplastic tissues, we emphasized genes that molecularly distinguish benign tissue from malignant—possibly accounting for the abundance of cancer-related genes picked up by our screen. Combining tissue microarrays, cDNA microarrays, and linked clinical and pathology data will facilitate rapid characterization of candidate biomarkers and regulatory cancer genes, as demonstrated here with respect to hepsin and PIM1 kinase. □

Methods

Preparation of total RNA and reference pools

The prostate surgical specimens were obtained from The University of Michigan Specialized Research Program in Prostate Cancer (SPORC) tumour bank with Institutional Review Board approval. Tissues were homogenized in Trizol (Gibco) and the total RNA was isolated according to the standard Trizol protocol. Total RNA integrity was judged by denaturing-formaldehyde agarose gel electrophoresis. Total RNA from four normal tissues was combined in equal concentrations to obtain the reference pool. The human prostate total RNA used in the commercial reference pool was obtained from Clontech.

Microarray procedures

Complementary DNA microarray analysis of gene expression was done essentially as described (available at <http://www.microarrays.org>). The sequence-verified cDNA clones on the human cDNA microarray are listed in the Supplementary Information and are available from the Research Genetics web site (<http://www.resgen.com>). Based on the latest build of Unigene, our 10K human cDNA microarray covers about 5,520 known, named genes and 4,464 ESTs. Protocols for printing and post-processing of arrays are available (<http://www.microarrays.org/protocols.html>).

Data analysis

Primary analysis was done with the Genepix software package. Cy3-to-Cy5 ratios are determined for the individual genes along with various other quality-control parameters (for example, intensity over local background). Furthermore, bad spots or areas of the array with obvious defects were manually flagged. Flagged spots were not included in subsequent analyses.

These files were then imported into a Microsoft Access database. Before clustering, the normalized median of ratio values of the genes were log₂ transformed, filtered for presence across arrays, and selected for expression levels and patterns depending on the experimental set as stated in the figure legends. Average-linkage hierarchical clustering of an uncentred Pearson correlation similarity matrix was applied with the program Cluster¹¹, and the results were analysed; figures were generated with the program TreeView. Before hierarchical clustering, the data were filtered for at least a twofold change in expression ratio and ratio measurements present in 50% of the samples, thus yielding 1,520 genes from the NAP data set (Fig. 1b). Similarly, a threefold change with 75% measurements present yielded 1,006 genes from the commercial pool (Fig. 1c) data set.

Northern blot analysis

Thirty micrograms of total RNA were resolved by denaturing-formaldehyde agarose gel and transferred onto Hybond membrane (Amersham). The signal was visualized and quantified by phosphorimaging. For relative fold estimation, the ratio between the signals obtained from hepsin and GAPDH probes was calculated. For northern analysis, we compared the intensity of the transcript (as determined by phosphorimaging) in NAP tissue with that of localized prostate cancer. The results are first normalized against a GAPDH control and a ratio of cancer to normal is then obtained. Similarly, for microarray analysis, the normalized Cy5/Cy3 quotient of normal tissue is compared to the Cy5/Cy3 quotient of prostate cancer and a ratio of cancer to normal tissue is obtained.

Immunohistochemistry and tissue microarrays

High-density tissue microarrays were assembled as previously described^{14,15}. Initial sections were stained for haematoxylin and eosin to verify histology. Standard biotin-avidin-complex immunohistochemistry was performed. The affinity-purified polyclonal rabbit antibody against hepsin was used at a 1:40 dilution (original concentration 0.2 mg ml⁻¹) for this study. Immunostaining intensity was scored by a genito-urinary pathologist (M.A.R.) as absent, weak, moderate or strong. Scoring was performed blind using a telepathology system without knowledge of overall Gleason score (for example, tumour grade), tumour size or clinical outcome¹⁵. A total of 738 tissue samples from benign ($n = 205$), HG-PIN ($n = 38$), localized prostate cancer ($n = 335$) and metastatic prostate cancer ($n = 160$) were examined.

Similarly, PIM1 was analysed using two tissue-microarray blocks from 810 prostate-cancer samples from 135 patients. Six prostate-cancer samples were evaluated from each case and a median score was calculated. In addition, a few samples with benign prostatic tissues (for example, benign epithelium and atrophy) and HG-PIN were examined. Immunohistochemistry was performed as above, using a rabbit polyclonal antibody against the carboxy terminus of PIM1 (Santa Cruz Biotechnology) at a dilution of 1:100.

Case selection

Cases of clinically localized prostate cancer were identified from the University of Michigan Prostate SPORC tumour bank. The advanced prostate tumours were collected from a series of rapid autopsies performed at the University of Michigan on men who died of metastatic prostate cancer. Autopsies were performed 4–6 h after death. The clinical and pathologic findings of these cases have recently been reported²⁰.

Statistical methods

A nonparametric analysis of variance test (Mann–Whitney, two categories) was used to evaluate whether the prostate samples expressed hepsin and PIM1 at different levels based on various parameters (tissue type, Gleason score and tumour size). Kaplan–Meier analysis was used to estimate the cumulative percentage of PSA-free progression ('survival'). The log-rank test was used to assess the differences in disease-free progression hepsin immunostaining. Cox proportional-hazard regression was used for multivariate analyses.

Received 16 January; accepted 26 June 2001.

- Abate-Shen, C. & Shen, M. M. Molecular genetics of prostate cancer. *Genes Dev.* **14**, 2410–2434 (2000).
- Ruijter, E. *et al.* Molecular genetics and epidemiology of prostate carcinoma. *Endocr. Rev.* **20**, 22–45 (1999).
- Barry, M. J. Prostate-specific-antigen testing for early diagnosis of prostate cancer. *N. Engl. J. Med.* **344**, 1373–1377 (2001).
- Liotta, L. & Petricoin, E. Molecular profiling of human cancer. *Nature Rev. Genet.* **1**, 48–56 (2000).
- Elek, J., Park, K. H. & Narayanan, R. Microarray-based expression profiling in prostate tumors. *In Vivo* **14**, 173–182 (2000).
- Emmert-Buck, M. R. *et al.* Molecular profiling of clinical tissue specimens: feasibility and applications. *Am. J. Pathol.* **156**, 1109–1115 (2000).
- Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
- Bittner, M. *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540 (2000).
- Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
- Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
- Tomita, K. *et al.* Cadherin switching in human prostate cancer progression. *Cancer Res.* **60**, 3650–3654 (2000).
- Shurbaji, M. S., Kalbfleisch, J. H. & Thurmond, T. S. Immunohistochemical detection of a fatty acid synthase (OA-519) as a predictor of progression of prostate cancer. *Hum. Pathol.* **27**, 917–921 (1996).
- Nononen, J. *et al.* Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Med.* **4**, 844–847 (1998).
- Perrone, E. E. *et al.* Tissue microarray assessment of prostate cancer tumor proliferation in African-American and white men. *J. Natl Cancer Inst.* **92**, 937–939 (2000).
- Tsuji, A. *et al.* Hepsin, a cell membrane-associated protease. Characterization, tissue distribution, and gene localization. *J. Biol. Chem.* **266**, 16948–16953 (1991).
- Pound, C. R. *et al.* Natural history of progression after PSA elevation following radical prostatectomy. *Jama* **281**, 1591–1597 (1999).

- Gleason, D. F. Histologic grading of prostate cancer: a perspective. *Hum. Pathol.* **23**, 273–279 (1992).
- Buttayan, R., Sawczuk, I. S., Benson, M. C., Siegal, J. D. & Olsson, C. A. Enhanced expression of the c-myc protooncogene in high-grade human prostate cancers. *Prostate* **11**, 327–337 (1987).
- Rubin, M. A. *et al.* Rapid ('warm') autopsy study for procurement of metastatic prostate cancer. *Clin. Cancer Res.* **6**, 1036–10345 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We are indebted to P. Ward and the University of Michigan Microarray Network for support and encouragement in the development of the Pathology Microarray Node. We thank M. Sanda and J. Wei for support of the clinical database; E. Cushenberry for her help with collecting the prostate tissue samples, The University of Michigan Comprehensive Cancer Center Histology and Immunoperoxidase Core; and the instructors of the 1999 Cold Spring Harbor Workshop on DNA Microarrays. This work is supported by developmental grants (A.M.C. and M.A.R.) from the Specialized Program of Research Excellence in Prostate Cancer, National Cancer Institute.

Correspondence and requests for materials should be addressed to A.M.C. (e-mail: arul@umich.edu).

Haematopoietic cell-specific CDM family protein DOCK2 is essential for lymphocyte migration

Yoshinori Fukui*, Osamu Hashimoto*†, Terukazu Sanui*, Takamasa Oono*, Hironori Koga†, Masaaki Abe‡, Ayumi Inayoshi*, Mayuko Noda*, Masahiro Oike§, Toshikazu Shirai‡ & Takehiko Sasazuki*||

* Division of Immunogenetics, Department of Immunobiology and Neuroscience, Medical Institute of Bioregulation, Kyushu University, and CREST, Japan Science and Technology Corporation, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan

† Research Center for Innovative Cancer Therapy, Kurume University School of Medicine, Asahimachi 67, Kurume 830-0011, Japan

‡ The Second Department of Pathology, Juntendo University School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

§ Department of Pharmacology, Graduate School of Medical Sciences, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan

|| International Medical Center of Japan, 1-12-1 Toyama, Shinjuku-ku, Tokyo 162-8655, Japan

Cell migration is a fundamental biological process involving membrane polarization and cytoskeletal dynamics¹, both of which are regulated by Rho family GTPases^{2–5}. Among these molecules, Rac is crucial for generating the actin-rich lamellipodial protrusion, a principal part of the driving force for movement^{3,6}. The CDM family proteins, *Caenorhabditis elegans* CED-5, human DOCK180 and *Drosophila melanogaster* Myoblast City (MBC), are implicated to mediate membrane extension by functioning upstream of Rac^{7–12}. Although genetic analysis has shown that CED-5 and Myoblast City are crucial for migration of particular types of cells^{8,9,12}, physiological relevance of the CDM family proteins in mammals remains unknown. Here we show that DOCK2, a haematopoietic cell-specific CDM family protein¹³, is indispensable for lymphocyte chemotaxis. DOCK2-deficient mice (DOCK2^{-/-}) exhibited migration defects of T and B lymphocytes, but not of monocytes, in response to chemokines, resulting in several abnormalities including T lymphocytopenia, atrophy of lymphoid follicles and loss of marginal-zone B cells. In DOCK2^{-/-} lymphocytes, chemokine-induced Rac activation and actin poly-

merization were almost totally abolished. Thus, in lymphocyte migration DOCK2 functions as a central regulator that mediates cytoskeletal reorganization through Rac activation.

DOCK2, originally described as KIAA0209 (ref. 14), encodes another member of the CDM family proteins, which is specifically expressed in human haematopoietic cells¹³. Although this protein has been shown to bind to and activate Rac in 293T kidney cells¹³, its physiological function remains unknown. In an attempt to search the genes that are expressed in the mouse thymus, we isolated complementary DNA, designated as *Hch* (GenBank accession number AYO27438), which predictably encodes 1,828 amino acids including the Src-homology (SH)-3 domain at the amino terminus. *Hch* was homologous to CED-5, MBC and DOCK180; furthermore, 1,677 of the 1,828 amino acids of *Hch* were identical to human DOCK2. Whereas *DOCK180* was expressed in various

tissues, the expression of *Hch* was restricted to thymus, spleen and lymph nodes (data not shown), suggesting that *Hch* is predominantly expressed in haematopoietic cells. The *Hch* expression was detected in all T-, B- and monocyte cell lines tested, with the exception of two T-cell lines, BW5147 α β ⁻ and its derivative BE α 16-3 (16-3; data not shown). From these results, we conclude that *Hch* is a mouse homologue of human DOCK2.

We first examined whether the expression of DOCK2 affects the actin cytoskeleton by introducing mouse DOCK2 into a T-cell line, 16-3, lacking endogenous expression. In western blot analysis two stable transfectants, 17-11 and 25-7, yielded specific bands showing slightly higher (17-11) or lower (25-7) expression compared with a wild-type cell line N3-5 (Fig. 1a). The GTP-bound, activated Rac, which interacts with PAK1 (ref. 15), was detected with 17-11 and N3-5, but not with 16-3 (Fig. 1b). In several aspects, we found that

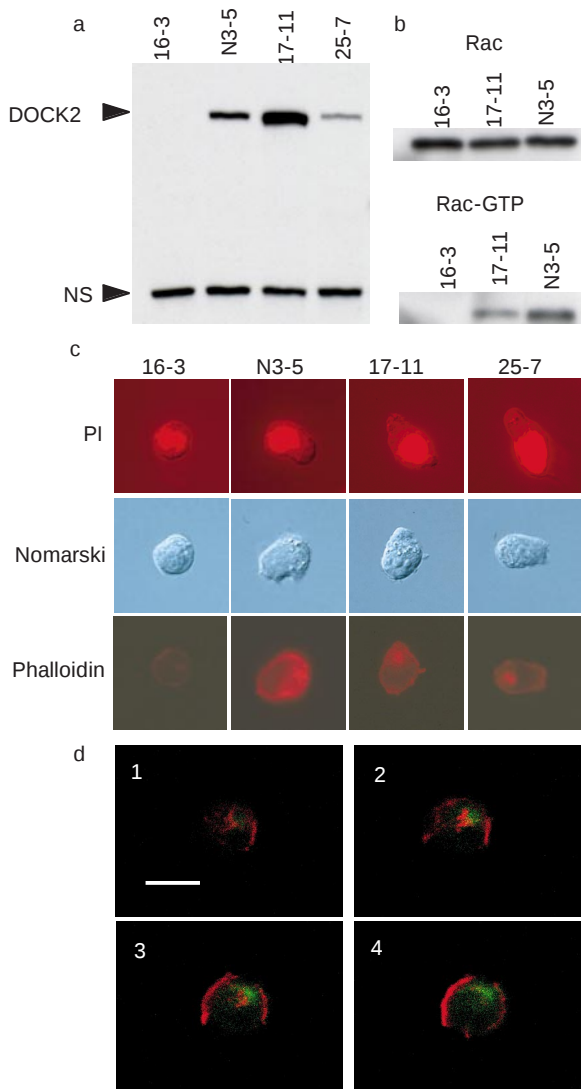


Figure 1 Remodelling of actin cytoskeleton by DOCK2 in a T-cell line. **a**, Analysis of total cell extracts by immunoblotting using anti-DOCK2 polyclonal antibody. A non-specific band (NS) is included as a loading control. **b**, Analysis of total cell extracts using anti-Rac monoclonal antibody with (bottom panel) or without (top panel) incubation of GST-fusion PAK1 RBD. **c**, Cells were stained with propidium iodide (PI) or phalloidin and analysed using a fluorescence microscope with Nomarski optics. Nomarski- and phalloidin-stained images were taken from the same cell. **d**, After staining 25-7 cells with anti-HA antibody for detection of DOCK2 (green) and phalloidin (red), the cell images were taken with a laser confocal microscope by scanning the cells from the top to the bottom (from 1 to 9) by a 2- μ m interval. Scale bar, 1 μ m.

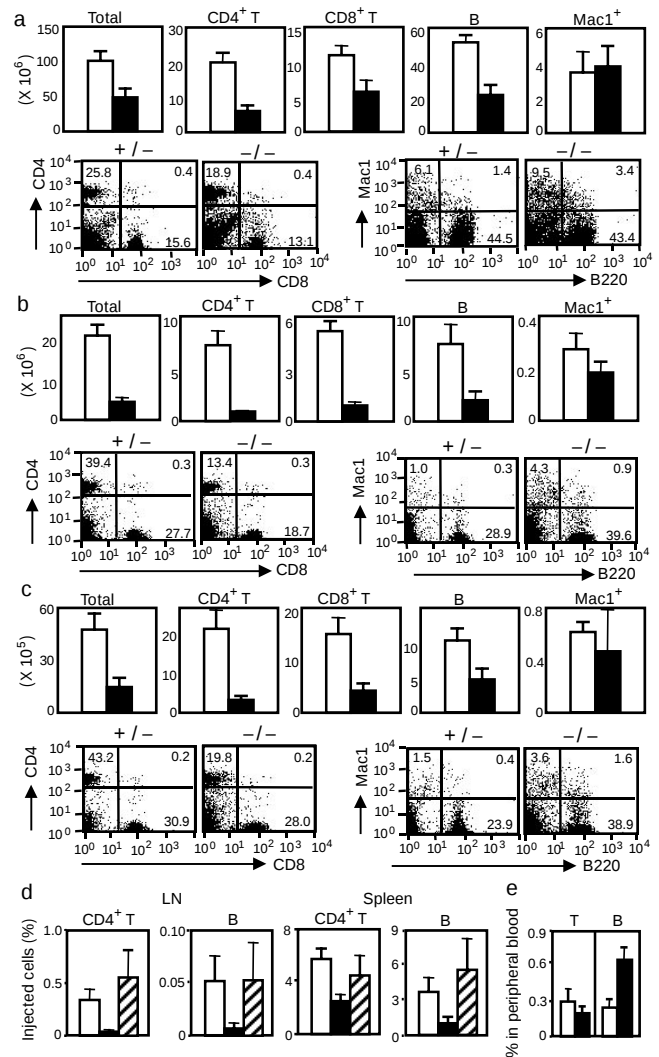


Figure 2 Impaired homing of DOCK2^{-/-} lymphocytes into secondary lymphoid organs. **a–c**, Spleen and lymph node cells were stained for CD4, CD8, B220 and Mac1. The number of total and each subset of cells in spleen (**a**), mesenteric lymph nodes (**b**) and inguinal lymph nodes (**c**) are compared between DOCK2^{+/+} (open bars; $n = 4$) and DOCK2^{-/-} (filled bars; $n = 7$ in **a** and 4 in **b** and **c**). **d**, Analysis of inguinal lymph nodes and spleen for migration of the fluorescence-labelled CD4⁺ T and B cells at 48 h after transfer. Open bars, from DOCK2^{+/+} to C57BL/6 ($n = 4$); filled bars, from DOCK2^{-/-} to C57BL/6 ($n = 5$); hatched bars, from DOCK2^{+/+} to DOCK2^{-/-} ($n = 5$). **e**, Analysis of peripheral blood T or B cells for percentages of the transferred cells at 48 h after transfer. Open bar, from DOCK2^{+/+} to C57BL/6 ($n = 7$ for T cells; $n = 5$ for B cells); filled bar, from DOCK2^{-/-} to C57BL/6 ($n = 6$).

cell lines 17-11 and 25-7 were morphologically different from 16-3, but were similar to N3-5. Although 16-3 had a round cell shape, 17-11, 25-7 and N3-5 showed a relatively flat and polarized cell shape with a biased localization of the nucleus (Fig. 1c). When cells were stained with phalloidin, F-actin assembly was observed in 17-11, 25-7 and N3-5, but not in 16-3 (Fig. 1c). The confocal laser microscopic analysis revealed that the introduced DOCK2 did not co-localize with, but was located adjacent to, this F-actin assembly at the root of the lamellipodial protrusion (Fig. 1d). Thus, like other CDM family proteins, DOCK2 mediates remodelling of actin cytoskeleton in T cells through Rac activation.

To investigate the physiological role of DOCK2, we generated DOCK2-deficient mice by homologous recombination in embryonic stem (ES) cells. DOCK2^{-/-} mice were born at the expected mendelian ratio without apparent abnormality. However, the total number of spleen cells in DOCK2^{-/-} mice was reduced to about 50% of that in DOCK2^{+/+} mice (Fig. 2a). Although the proportions of CD4⁺ and CD8⁺ T cells, and also B cells, were not much changed, the proportion of Mac1⁺ monocytes significantly increased in DOCK2^{-/-} mice. A similar tendency was observed with

mesenteric and inguinal lymph node cells, except for the proportion of CD4⁺ T cells (Fig. 2b, c): whereas CD4⁺ T cells occupied approximately 40% of lymph node cells in DOCK2^{+/+} mice, this proportion decreased to less than 25% in DOCK2^{-/-} mice. Collectively, the amounts of T and B cells, but not monocytes, were markedly reduced in the spleen and lymph nodes of DOCK2^{-/-} mice.

This finding raises the possibility that lymphocyte homing to secondary lymphoid organs might be impaired in DOCK2^{-/-} mice. To address this issue, we compared the migratory capacity of fluorescence-labelled CD4⁺ T or B cells to the peripheral lymph nodes or spleen. When cells prepared from DOCK2^{+/+} mice were injected intravenously into wild-type mice, 0.34 and 0.05% of the injected CD4⁺ T or B cells, on an average, migrated to the inguinal lymph nodes, respectively (Fig. 2d, left). However, with cells prepared from DOCK2^{-/-} mice, the percentages of the migrated CD4⁺ T and B cells were reduced to about 10% of the level observed in DOCK2^{+/+} mice. A similar, but less marked difference was found on analysis of migration to the wild-type spleen (Fig. 2d, right). On the other hand, CD4⁺ T and B cells prepared from DOCK2^{+/+} mice efficiently migrated to the inguinal lymph nodes and spleen of

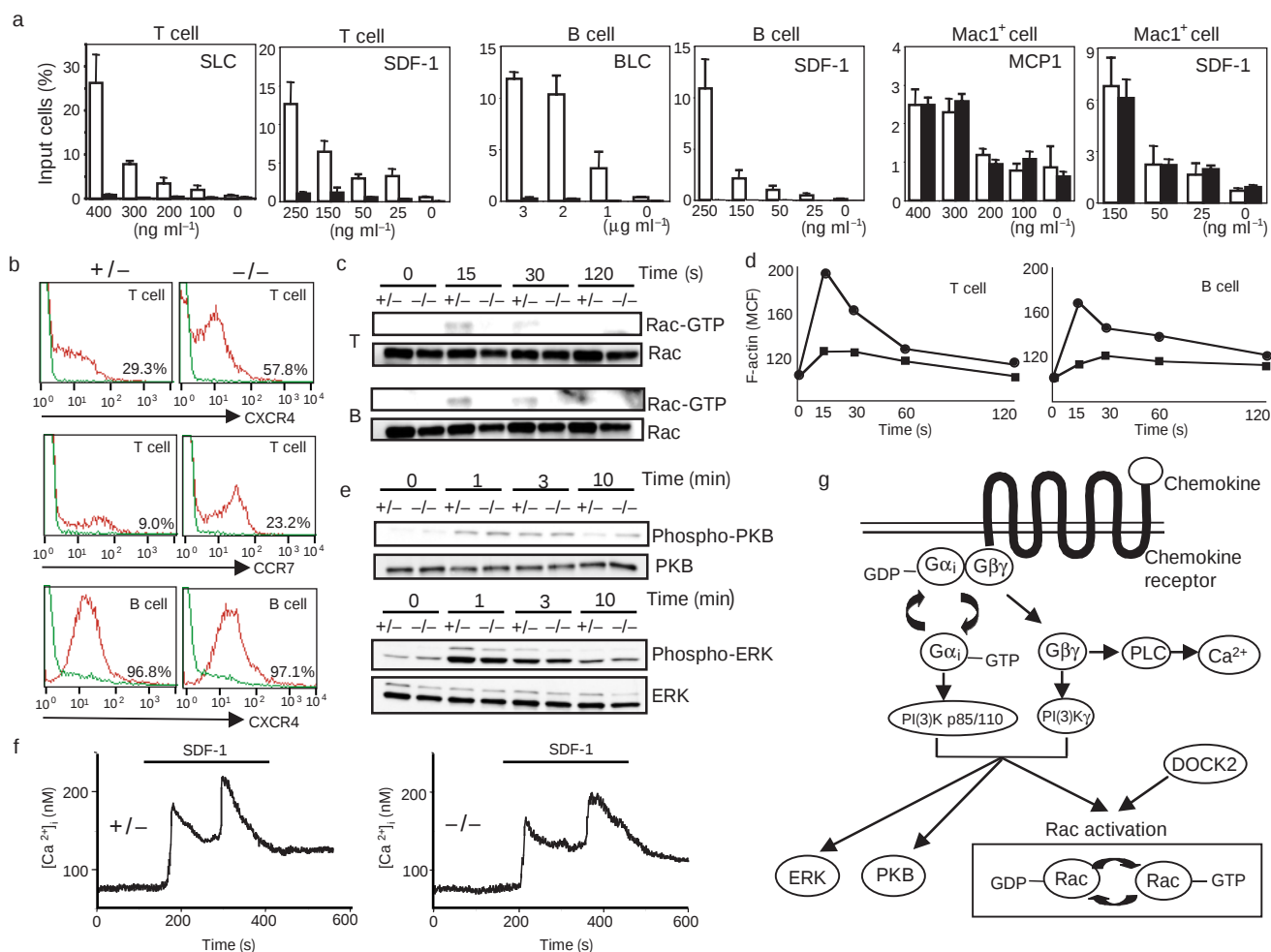


Figure 3 Lymphocyte migration defect in DOCK2^{-/-} mice. **a**, Spleen cells of DOCK2^{+/+} (open bar) and DOCK2^{-/-} (filled bar) were used in transwell chemotaxis assay. Results are expressed as the percentages of the input cells. **b**, Spleen cells were stained with anti-CXCR4 or anti-CCR7 antibody³⁰ (red) or control antibody (green) in combination with monoclonal antibodies against Thy1.2 or B220, and Thy1.2⁺ T cells or B220⁺ B cells were analysed for receptor expression (graphs show relative cell numbers). **c**, Splenic T and B cells (1.5 × 10⁷) treated with SDF-1 (500 ng ml⁻¹) for the indicated times were analysed for Rac activation using GST-fusion PAK1 RBD. **d**, Splenic T and B cells (5 × 10⁶) from DOCK2^{-/-} (squares) and DOCK2^{+/+} (circles) mice were treated with SDF-1 (500 ng ml⁻¹) for the indicated times and analysed for actin polymerization with flow cytometry. Results

are expressed as the average of mean channel fluorescence (MCF) in three independent experiments, with the baseline fluorescence arbitrarily assigned a value of 100. **e**, Splenic B cells (5–8 × 10⁶) treated with SDF-1 (500 ng ml⁻¹) for the indicated times were analysed for phosphorylation of PKB or ERK using the phospho-specific antibody against PKB (Ser⁴⁷³) or ERK1 and ERK2 (Thr²⁰²/Tyr²⁰⁴), respectively. **f**, Fura-2-loaded splenic B cells were stimulated with SDF-1 at 250 ng ml⁻¹, and [Ca²⁺]_i was monitored with a digital fluorescence microscopy system. **g**, Schematic representation for signalling pathways from chemokine receptor to Rac activation. For simplicity, the potential intermediates in the signalling pathways are not shown.

DOCK2^{-/-} mice (Fig. 2d). These results suggest that lymphocyte homing is impaired in DOCK2^{-/-} mice owing to an intrinsic defect in lymphocytes. Although no significant difference was found in the frequency of transferred T cells in the peripheral blood of wild-type mice, the percentage of transferred DOCK2^{-/-} B cells showed a 2.7-fold increase in level compared with DOCK2^{+/-} B cells (Fig. 2e).

We next analysed chemotaxis of splenic T and B cells in response to several chemokines, such as secondary lymphoid tissue chemokine (SLC), B-lymphocyte chemoattractant (BLC), stroma-derived factor (SDF)-1 and monocyte chemoattractant protein (MCP)-1. Whereas T and B cells from DOCK2^{+/-} mice efficiently migrated in response to SLC, SDF-1 and BLC, those from DOCK2^{-/-} mice did not show any response to these chemokines (Fig. 3a, left and middle). In contrast, the chemotactic responses of DOCK2^{-/-} monocytes to MCP-1 and SDF-1 were as efficient as the responses of DOCK2^{+/-} mice (Fig. 3a, right). The chemotactic activities of SLC

and SDF-1 are mediated through interaction with their receptors expressed on target cells, CCR7 or CXCR4, respectively¹⁶⁻¹⁸. Splenic B cells from DOCK2^{-/-} and DOCK2^{+/-} mice comparably expressed CXCR4 on their surface, and the expression of CXCR4 and CCR7 on splenic T cells was even higher in DOCK2^{-/-} than in DOCK2^{+/-} mice (Fig. 3b). Thus, it is clear that the defect in chemotactic response of DOCK2^{-/-} lymphocytes does not result from extremely low, or absence of, receptor expression. Taken together, these results indicate that although DOCK2 is dispensable for monocyte migration, this molecule has a crucial role in lymphocyte migration by functioning downstream of chemokine receptors.

Chemokine receptors are coupled with heterodimeric G_i proteins that activate a variety of signalling pathways. Phosphatidylinositol-3-OH kinases (PI(3)K) are some of the central signalling molecules thought to be involved in activation of Rac, protein kinase B (PKB) and extracellular signal-regulated kinases (ERKs)^{19,20}. When T cells

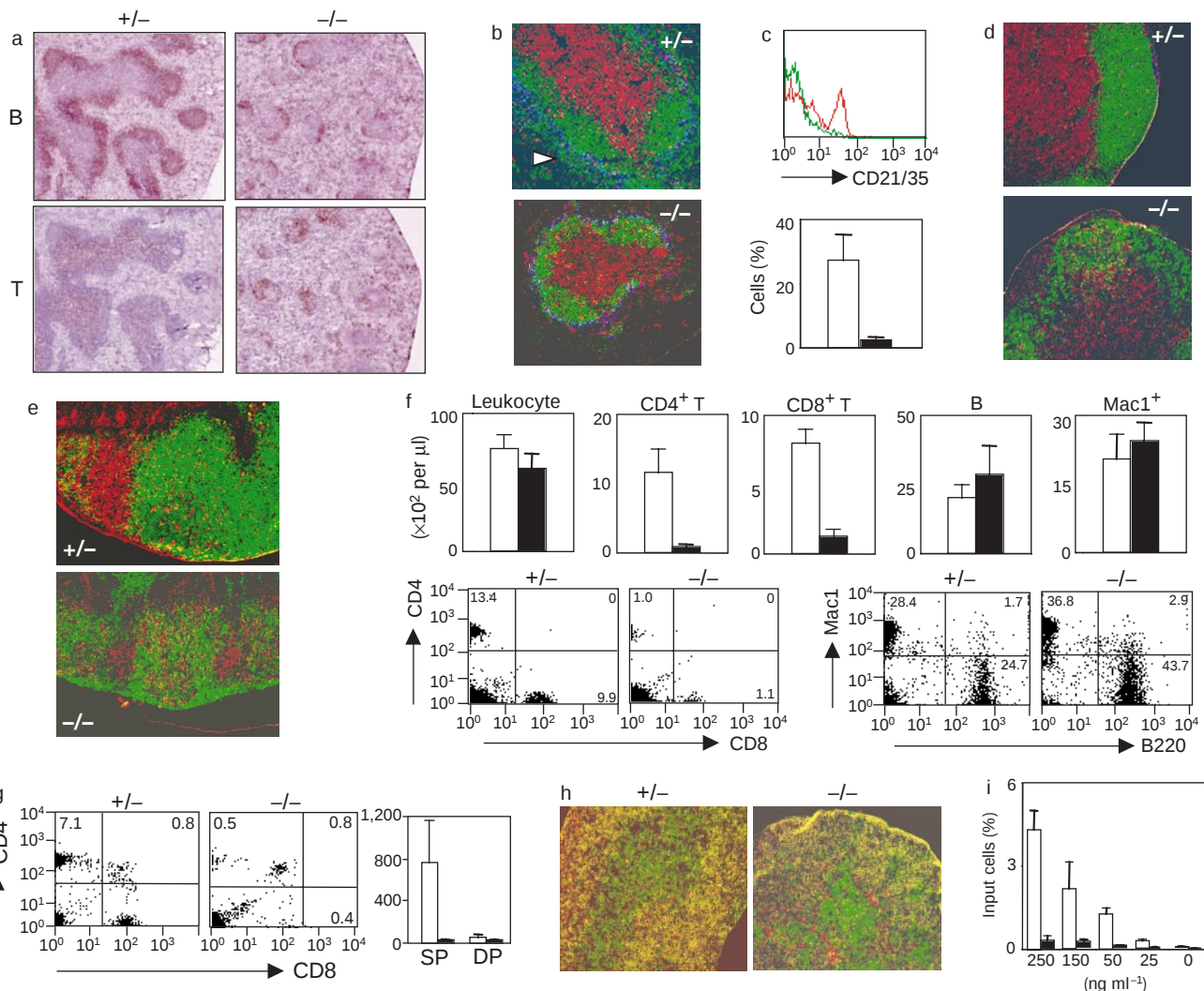


Figure 4 Abnormal architecture of the immune system in DOCK2^{-/-} mice. **a**, Staining of serial spleen sections for B cells (top panels) or T cells (bottom panels) with anti-B220 or anti-Thy1.2 monoclonal antibody, respectively. **b**, Spleen sections were stained for B cells (green), T cells (red) and metallophilic macrophages (blue) with anti-B220 monoclonal antibody, a mixture of anti-CD4 and anti-CD8 monoclonal antibodies, and MOMA1. The arrowhead indicates marginal-zone B cells. **c**, Spleen cells were stained for B220, CD23 and CD21/35, and B220⁺CD23⁻ B cells were analysed for the expression of CD21/35 (graph shows relative cell numbers). The representative flow cytometric profiles and the percentages of CD21/35⁺ cells in the gated population of DOCK2^{+/-} (red, open bar) and DOCK2^{-/-} (green, filled bar) spleen cells are shown. **d**, Lymph node sections were stained for B cells (green) and T cells (red) as in **b**. **e**, Peyer's patches were stained for B cells

(green) and T cells (red) as in **b**. **f**, Blood samples were counted and stained for CD4, CD8, B220 and Mac1. The number of total and each subset of peripheral blood leukocytes are compared between DOCK2^{+/-} (open bars; *n* = 4) and DOCK2^{-/-} (filled bars; *n* = 4). **g**, Thymus organ cultures were used for transwell chemotaxis assays in the presence of ELC. Left, expression of CD4 and CD8 on the emigrated thymocytes; right, number of CD4⁺CD8⁻ or CD4⁺CD8⁺ mature thymocytes (SP) and CD4⁺CD8⁺ thymocytes (DP) collected for 90 s from DOCK2^{+/-} (open bar; *n* = 3) and DOCK2^{-/-} (filled bar; *n* = 3) thymus. **h**, Thymus sections stained for CD4 (green) and CD8 (red). **i**, Chemotactic response of CD4⁺CD8⁺ thymocytes to SDF-1 was compared between DOCK2^{+/-} (open bars) and DOCK2^{-/-} (filled bars) using a transwell chemotaxis assay. Results are expressed as percentages of the input cells.

and B cells from DOCK2^{+/-} mice were stimulated with SDF-1, the activated Rac was detected. However, such activation was totally abolished in DOCK2^{-/-} lymphocytes (Fig. 3c). Consistent with this result, SDF-1-induced actin polymerization was observed in DOCK2^{+/-} lymphocytes, but was scarcely detected in DOCK2^{-/-} lymphocytes (Fig. 3d). In contrast, SDF-1-induced phosphorylations of PKB, ERK1 and ERK2 were comparable between DOCK2^{+/-} and DOCK2^{-/-} B cells (Fig. 3e). Chemokines can also activate phospholipase C leading to an increase in intracellular Ca²⁺ concentration [Ca²⁺]_i (ref. 20). Mobilization of Ca²⁺ induced by SDF-1 occurred to a similar extent in both DOCK2^{+/-} and DOCK2^{-/-} B cells (Fig. 3f). These results indicate that in chemokine receptor-mediated signalling pathways DOCK2 is predominantly involved in Rac activation (Fig. 3g).

The chemokines SLC and BLC are critical for homeostatic trafficking of lymphocytes into subcompartments of lymphoid organs²¹⁻²⁴. As DOCK2^{-/-} lymphocytes lacked chemotactic response to these chemokines *in vitro*, we made a close examination of spleen, lymph nodes and Peyer's patches. In the spleen of DOCK2^{+/-} mice, lymphoid follicles that consisted of T-cell zone and the surrounding B-cell area were clearly defined (Fig. 4a, left). However, such lymphoid follicles were quite atrophic in DOCK2^{-/-} spleen, and scattered distribution of T and B cells in red pulp was observed in this line (Fig. 4a, right). Immunofluorescence analysis of splenic tissue sections revealed that marginal-zone B cells, normally distributed around metallophilic macrophages, were markedly reduced in DOCK2^{-/-} mice (Fig. 4b). The marginal-zone B cells are characterized as CD21/35^{high}CD23^{neg-low} B cells²⁵. Although around 25–30% of B220⁺CD23⁻ splenic B cells expressed CD21/CD35 in DOCK2^{+/-} mice, this population of B cells was scarcely found in DOCK2^{-/-} mice, confirming the defect of marginal-zone B cells in this line (Fig. 4c). The atrophy of lymphoid follicles and aberrant T-cell distribution were also observed in lymph nodes of DOCK2^{-/-} mice (Fig. 4d). In addition, we found that Peyer's patches were poorly developed in DOCK2^{-/-} mice in terms of size and cell density (Fig. 4e).

SDF-1 serves as a chemoattractant for immature lymphocytes. The chemotactic responses of DOCK2^{-/-} pro-B and pre-B cells to SDF-1 (250 ng ml⁻¹) were significantly reduced compared with those of DOCK2^{+/-} mice (pro-B cells, 3.4% versus 18.5%; pre-B cells, 0.5% versus 5.6%, respectively). However, the amounts of pro-B cells, pre-B cells, immature B cells and myeloid cells in the bone marrow were unchanged between DOCK2^{+/-} and DOCK2^{-/-} mice (data not shown), indicating that DOCK2 deficiency affects neither B lymphopoiesis nor myelopoiesis. Although the total number of thymocytes and the proportion of CD4⁺CD8⁻ thymocytes in DOCK2^{-/-} mice decreased and increased, respectively, compared with those in DOCK2^{+/-} mice, considerable numbers of mature thymocytes were still observed in these lines (data not shown). Nonetheless, the amounts of CD4⁺ and CD8⁺ T cells in peripheral blood decreased extremely in DOCK2^{-/-} mice (Fig. 4f). We compared thymocyte emigration in response to EBI1-ligand chemokine (ELC), which serves as a chemoattractant for mature thymocytes²⁶. Although CD4⁺CD8⁻ and CD4⁺CD8⁺ mature thymocytes efficiently emigrated from the DOCK2^{+/-} thymus, mature thymocytes were scarcely detected in the case of DOCK2^{-/-} mice (Fig. 4g, left). The efficacy of mature thymocyte emigration from DOCK2^{-/-} thymus was reduced to less than 5% of the wild-type level (Fig. 4g, right). It is thus suggested that an emigration defect of mature thymocytes to the peripheral bloodstream is involved in T lymphocytopenia observed in DOCK2^{-/-} mice. We found that in the thymus of DOCK2^{-/-} mice, CD4⁺CD8⁻ or CD4⁺CD8⁺ mature thymocytes were distributed irregularly throughout the thymus as small patches (Fig. 4h). Although the precise mechanism for this remains unknown, thymocyte trafficking within the thymus may also be impaired in DOCK2^{-/-} mice. Supporting this, the chemotactic response of CD4⁺CD8⁺

thymocytes to SDF-1 was severely impaired in DOCK2^{-/-} mice (Fig. 4i).

Our results provide genetic, functional and biochemical evidence that haematopoietic cell-specific CDM family protein DOCK2 functions as a central molecule that activates Rac and mediates cytoskeletal reorganization during lymphocyte migration. Consistent with the unresponsiveness to 'homeostatic chemokines', several abnormalities that were observed in DOCK2^{-/-} mice were similar, although not identical, to those of mice that lacked CCR7, SLC or CXCR5 (the receptor for BLC)²¹⁻²⁴. However, DOCK2^{-/-} mice exhibited additional phenotypes such as T lymphocytopenia, loss of marginal-zone B cells, abnormal thymus architecture and reduced lymphocyte homing to the spleen. Some of these features probably reflect unresponsiveness of DOCK2^{-/-} lymphocytes to other known or undefined chemokines. Furthermore, because of its nature to regulate the actin cytoskeleton, DOCK2 deficiency may involve more than just a failure of chemokine-mediated lymphocyte migration.

In spite of its critical role in lymphocyte migration, DOCK2 deficiency does not affect the chemotactic response of monocytes. This is in contrast with results on mice that lack PI(3)Kγ, where migrations of neutrophils and monocytes are impaired through Rac-independent mechanisms²⁷⁻²⁹. In these mice, lymphocyte migration seems normal because the amount of T and B lymphocytes in peripheral blood and secondary lymphoid organs was similar to those of wild-type mice²⁷. Our results, together with these previous reports, suggest that migration of lymphocytes and myeloid cells are regulated through distinct signalling pathways. DOCK2 has been shown to be expressed in tissue macrophages in humans¹³, and here its expression was confirmed in mouse monocyte cell lines. However, in western blot analysis DOCK2 expression was scarcely detected in both splenic and thioglycollate-induced peritoneal macrophages of wild-type mice. On the other hand, we have found in a recent study that DOCK180, originally described to be expressed in non-haematopoietic cells⁷, is expressed in monocytes but not in lymphocytes (data not shown). Therefore, given that the CDM family proteins are important for cell migration, monocyte migration may be mediated by DOCK180. Although DOCK180 is associated with CRK, which is implicated in signalling from focal adhesion⁷, DOCK2 does not bind to this protein¹³. This also raises the possibility that DOCK2 and DOCK180 may have evolved from a common ancestral gene to be in charge of distinct functions in lymphocytes and monocytes. At this stage it is unknown as to what extent DOCK2 and DOCK180 are functionally complementary, an issue that should be investigated in future studies. □

Methods

Stable transfectants

The gene encoding full-length DOCK2 with or without a haemagglutinin (HA) tag was created under a PB1 expression vector and introduced into the 16-3 T-cell line by electroporation.

Generation of DOCK2^{-/-} mice

Genomic clones containing the mouse *DOCK2* gene were isolated from a 129/Sv genomic library. The targeting vector was constructed to replace a 3.1-kilobase (kb) fragment from the *KpnI* site located within exon 3 to the *BamHI* site between exon 5 and 6 with a neomycin resistance cassette (*neo*). Correctly targeted ES clones were microinjected into C57BL/6 blastocysts, and the male chimaeras obtained were crossed with C57BL/6 female mice. Mice that were heterozygous for the mutant allele were crossed to obtain DOCK2^{-/-} mice. Southern blot analysis confirmed that wild-type *DOCK2* alleles were disrupted in mutant mice. Polymerase chain reaction with reverse transcription revealed that insertion of the *neo* cassette results in an appearance of a stop codon at the N-terminal portion of DOCK2, leaving only 42 amino-acid residues presumably intact. Supporting this, DOCK2 expression was not detected in the spleen and thymus of mutant mice in a western blot analysis. In all assays, DOCK2^{+/-} and DOCK2^{-/-} littermates were used.

Immunoblot analysis

Anti-mouse DOCK2 polyclonal antibody was developed by immunizing rabbits with keyhole limpet haemocyanin-conjugated, carboxy-terminal peptide, and this was used to

detect DOCK2 expression. To assess Rac activation, cell extracts were incubated with glutathione S-transferase (GST)-fusion, Rac-binding domain (RBD) of PAK1, and subjected to immunoblot analysis using anti-Rac monoclonal antibody (Upstate Biotechnology). Activation of PKB or ERK was assessed with phospho-specific antibodies against PKB (Ser⁴⁷³) or ERK1 and ERK2 (Thr²⁰²/Tyr²⁰⁴), respectively (Cell Signaling).

Immunofluorescent microscopy

Cells were fixed with 4% formaldehyde in phosphate-buffered saline (PBS), permeabilized with 0.1% saponin and stained with propidium iodide (CALBIOCHEM), Alexa 568-labelled phalloidin (Molecular Probes) and/or anti-HA antibody (Santa Cruz) followed by Alexa 488-labelled anti-rabbit immunoglobulin (Ig)- γ . The staining profiles were obtained with either a fluorescence microscope equipped with Nomarski optics or a confocal laser-scanning microscope equipped with an argon/krypton laser capable of dual excitation and detection.

Blood counts and flow cytometric analysis

Blood was obtained by retro-orbital venous plexus puncture and subjected to an automatic cell counter. Cell counts were determined for bone marrow cells, thymocytes, spleen cells and lymph node cells using Neubauer chambers. Flow cytometry was carried out by staining the cells with the relevant monoclonal antibodies (PharMingen) and analysing them on a FACScan (Becton-Dickinson). To assess actin polymerization, purified T and B cells were fixed with 4% paraformaldehyde in PBS, treated with 0.1% saponin, and stained with Alexa 488-labelled phalloidin for flow cytometry.

Tissue staining

Cryostat sections were fixed in acetone and 1% paraformaldehyde, and stained with Alexa 488- or 594-labelled monoclonal antibodies. Staining with anti-metallophilic macrophage monoclonal antibody (MOMA1) was followed by biotinylated anti-rat IgG antibody and developed with streptavidin-conjugated Cy5.5. In some experiments, tissue sections were stained with biotinylated monoclonal antibodies and visualized with a Vectastain ABC-PO kit (Vector Laboratories).

Lymphocyte homing assay

Purified CD4⁺ T or B cells (4×10^7 ml⁻¹) were labelled with 3 μ M BCECF-AM solution (Dojindo), washed with PBS and intravenously injected into mice (8×10^6 to 1×10^7 per mouse). After 48 h, cells were prepared from the spleen and inguinal lymph nodes, and counted. Before and after cell transfer, cells were stained with anti-CD4 or anti-B220 monoclonal antibody, and the percentages of migrated cells were calculated. The percentage of transferred T cells (4×10^6 per mouse) and B cells (1×10^7 per mouse) in peripheral blood was assessed by staining the cells with anti-Thy1.2 or anti-B220 monoclonal antibody at 48 h after transfer.

Chemotaxis assay

Thymus organ, spleen cells (1.5×10^6), bone marrow cells (1.5×10^6) and thymocytes (1.5×10^6) in 100 μ l RPMI medium were loaded into transwells (Coaster, 5- μ m pore size), which were placed onto 24-well plates containing 450 μ l RPMI medium supplemented with chemokines (Genzyme/Technique) at various concentrations. After 3 h of incubation at 37 °C, migrated cells to the lower chamber were collected and stained with the relevant antibodies. We carried out cell counts with a FACScan.

Measurement of intracellular Ca²⁺ concentration

We measured [Ca²⁺]_i using the Attofluor digital fluorescence microscopy system. Cells loaded with fura-2/AM (Dojindo) were put in a chamber of 0.5-ml volume and mounted on an inverted microscope. The bath was continuously perfused with a modified Krebs solution (pH 7.3). The fura-2 fluorescence images were recorded into a re-writable optical disc recorder at a rate of roughly 1 Hz and converted to apparent Ca²⁺ concentration.

Received 4 April; accepted 2 July 2001.

- Lauffenburger, D. A. & Horwitz, A. F. Cell migration: a physically integrated molecular process. *Cell* **84**, 359–369 (1996).
- Stowers, L., Yelon, D., Berg, L. J. & Chant, J. Regulation of the polarization of T cells toward antigen presenting cells by Ras-related GTPase Cdc42. *Proc. Natl Acad. Sci. USA* **92**, 5027–5031 (1995).
- Hall, A. Rho GTPases and the actin cytoskeleton. *Science* **279**, 509–514 (1998).
- Allen, W. E., Zicha, D., Ridley, A. J. & Jones, G. E. A role for Cdc42 in macrophage chemotaxis. *J. Cell Biol.* **141**, 1147–1157 (1998).
- Servant, G. *et al.* Polarization of chemoattractant receptor signaling during neutrophil chemotaxis. *Science* **287**, 1037–1040 (2000).
- Bar-Sagi, D. & Hall, A. Ras and Rho GTPases: a new family reunion. *Cell* **103**, 227–238 (2000).
- Hasegawa, H. *et al.* DOCK180, a major CRK-binding protein, alters cell morphology upon translocation to the cell membrane. *Mol. Cell. Biol.* **16**, 1770–1776 (1996).
- Erickson, M. R. S., Galletta, B. J. & Abmayr, S. M. *Drosophila myoblast city* encodes a conserved protein that is essential for myoblast fusion, dorsal closure, and cytoskeletal organization. *J. Cell Biol.* **138**, 589–603 (1997).
- Wu, Y.-C. & Horvitz, H. R. *C. elegans* phagocytosis and cell-migration protein CED-5 is similar to human DOCK180. *Nature* **392**, 501–504 (1998).
- Kiyokawa, E. *et al.* Activation of Rac1 by a Crk SH3-binding protein, DOCK180. *Genes Dev.* **12**, 3331–3336 (1998).
- Nolan, K. M. *et al.* Myoblast city, the *Drosophila* homolog of DOCK180/CED-5, is required in a Rac signaling pathway utilized for multiple developmental processes. *Genes Dev.* **12**, 3337–3342 (1998).

- Reddien, P. W. & Horvitz, H. R. CED-2/CrkII and CED-10/Rac control phagocytosis and cell migration in *Caenorhabditis elegans*. *Nature Cell Biol.* **2**, 131–136 (2000).
- Nishihara, H. *et al.* Non-adherent cell-specific expression of DOCK2, a member of the human CDM-family proteins. *Biochem. Biophys. Acta* **1452**, 179–187 (1999).
- Nagase, T. *et al.* Prediction of the coding sequences of unidentified human genes. VI. The coding sequences of 80 new genes (KIAA0201–KIAA0280) deduced by analysis of cDNA clones from cell line KG-1 and brain. *DNA Res.* **3**, 321–329.
- Manser, E., Leung, T., Salihuddin, H., Zhao, Z.-s. & Lim, L. A brain serine/threonine protein kinase activated by Cdc42 and Rac1. *Nature* **367**, 40–46 (1994).
- Bleul, C. C. *et al.* The lymphocyte chemoattractant SDF-1 is a ligand for LESTR/fusin and blocks HIV-1 entry. *Nature* **382**, 829–833 (1996).
- Oberlin, E. *et al.* The CXCR4 chemokine SDF-1 is the ligand for LESTR/fusin and prevents infection by T-cell-line-adapted HIV-1. *Nature* **382**, 833–835 (1996).
- Yoshida, R. *et al.* Secondary lymphoid-tissue chemokine is a functional ligand for the CC chemokine receptor CCR7. *J. Biol. Chem.* **273**, 7118–7122 (1998).
- Sotsios, Y. & Ward, S. G. Phosphoinositide 3-kinase: a key biochemical signal for cell migration in response to chemokines. *Immunol. Rev.* **177**, 217–235 (2000).
- Thelen, M. Dancing to the tune of chemokines. *Nature Immunol.* **2**, 129–134 (2001).
- Förster, R. *et al.* A putative chemokine receptor, BLR1, directs B cell migration to defined lymphoid organs and specific anatomic compartments of the spleen. *Cell* **87**, 1037–1047 (1996).
- Nakano, H. *et al.* A novel mutant gene involved in T-lymphocyte-specific homing into peripheral lymphoid organs on mouse chromosome 4. *Blood* **91**, 2886–2895 (1998).
- Förster, R. *et al.* CCR7 coordinates the primary immune response by establishing functional microenvironments in secondary lymphoid organs. *Cell* **99**, 23–33 (1999).
- Gunn, M. D. *et al.* Mice lacking expression of secondary lymphoid organ chemokine have defects in lymphocyte homing and dendritic cell localization. *J. Exp. Med.* **189**, 451–460 (1999).
- Oliver, A. M., Martin, F., Gartland, G. L., Carter, R. H. & Kearney, J. F. Marginal zone B cells exhibit unique activation, proliferative and immunoglobulin secretory responses. *Eur. J. Immunol.* **27**, 2366–2374 (1997).
- Kim, C. H., Pelus, L. M., White, J. R. & Broxmeyer, H. E. Differential chemotactic behavior of developing T cells in response to thymic chemokines. *Blood* **91**, 4434–4443 (1998).
- Sasaki, T. *et al.* Function of PI3K γ in thymocyte development, T cell activation, and neutrophil migration. *Science* **287**, 1040–1046 (2000).
- Li, Z. *et al.* Roles of PLC- β 2 and - β 3 and PI3K γ in chemoattractant-mediated signal transduction. *Science* **287**, 1046–1049 (2000).
- Hirsch, E. *et al.* Central role for G protein-coupled phosphoinositide 3-kinase γ in inflammation. *Science* **287**, 1049–1053 (2000).
- Onai, N. *et al.* Impairment of lymphopoiesis and myelopoiesis in mice reconstituted with bone marrow-hematopoietic progenitor cells expressing SDF-1-intrakinase. *Blood* **96**, 2074–2080 (2000).

Acknowledgements

We thank K. Matsushima and N. Onai for anti-CXCR4 and anti-CCR7 antibodies; K.-i. Nakayama and K. Nakayama for E14 ES cells and helpful suggestions on ES cell culture; P. Marrack for BE α 16-3; and Y. Yamada for blastocyst injection. This work was supported in part by a grant from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

Correspondence and requests for materials should be addressed to T.S. (e-mail: sasazuki@bioreg.kyushu-u.ac.jp).

The core of the motor domain determines the direction of myosin movement

Kazuaki Homma, Misako Yoshimura, Junya Saito, Reiko Ikebe & Mitsuo Ikebe

Department of Physiology, University of Massachusetts Medical School, 55 Lake Avenue North, Worcester, Massachusetts 01655-0127, USA

Myosins constitute a superfamily of at least 18 known classes of molecular motors that move along actin filaments^{1–3}. Myosins move towards the plus end of F-actin filaments; however, it was shown recently that a certain class of myosin, class VI myosin, moves towards the opposite end of F-actin, that is, in the minus direction⁴. As there is a large, unique insertion in the myosin VI head domain between the motor domain and the light-chain-binding domain (the lever arm), it was thought that this insertion alters the angle of the lever-arm switch movement, thereby changing the direction of motility⁴. Here we determine the

direction of motility of chimaeric myosins that comprise the motor domain and the lever-arm domain (containing an insert) from myosins that have movement in the opposite direction. The results show that the motor core domain, but neither the large insert nor the converter domain, determines the direction of myosin motility.

Until recently all characterized myosin motors have been shown to move towards the barbed end of actin filaments. As unconventional myosins are involved in diverse transport processes occurring along polarized actin cables, such as translocation of cellular organelles, a myosin having an opposite directionality of movement would be expected to have a unique cellular function. Of the 18

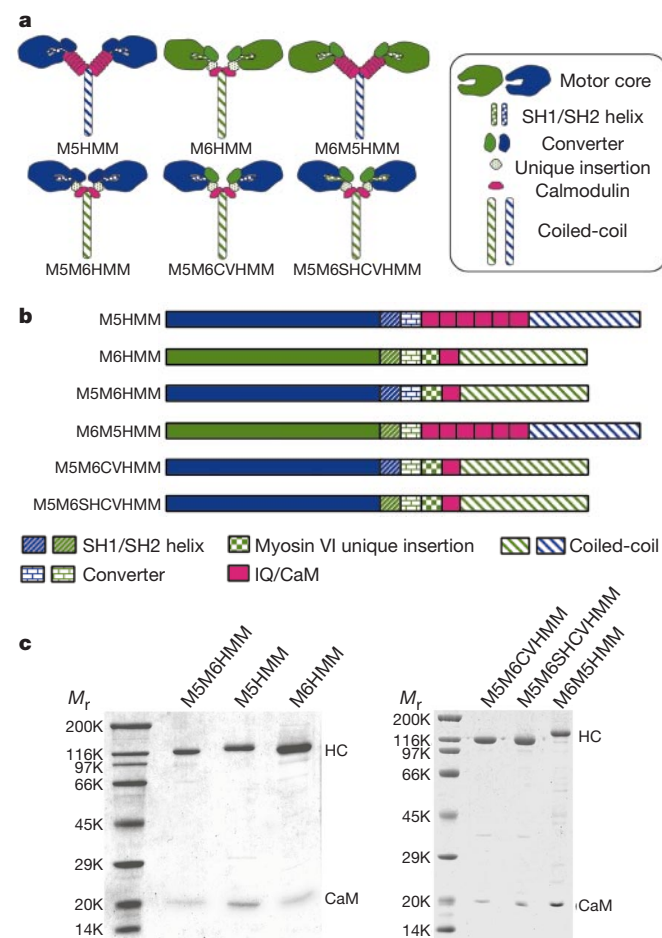


Figure 1 Structure of myosin constructs. **a**, Schematic representation of indicated myosin constructs. M5HMM consists of the two complete heads of myosin V (motor domain + six IQ motifs) plus the entire coiled-coil domain of myosin V. M6HMM consists of the two complete heads of myosin VI (motor domain + myosin-VI-specific large insert + one IQ motif) plus the entire coiled-coil domain of myosin VI. M5M6HMM is composed of two heads (containing the motor domain and converter domain of myosin V and the myosin-VI-specific large insert + myosin VI IQ motif) plus the entire coiled-coil domain of myosin VI. M5M6CVHMM substitutes the converter domain of M5M6HMM with the myosin VI converter domain. M5M6SHCVHMM substitutes the SH1/SH2 helix and converter domain of M5M6HMM with its myosin VI counterpart. M6M5HMM is composed of two heads (containing the motor domain and converter domain of myosin VI connected with the six myosin V IQ motifs, but without the large myosin-VI-specific insert) plus the entire coiled-coil domain of myosin V. **b**, Primary structure of various myosin constructs. Myosin V (blue) and myosin VI (green) sequences are indicated. Red box, IQ motif; chequered box, myosin-VI-specific large insertion between the converter and IQ domain; solid bar, the motor core. **c**, SDS-PAGE of purified myosin constructs. All purified constructs are composed of components with a high (heavy chain, HC) and low (light chain) molecular mass. The number of calmodulin light chains bound to each construct was estimated by determining the ratio of calmodulin light chain to heavy chain by densitometry.

different classes of myosins, it was found that class VI myosin moves towards the minus end of F-actin filaments⁴. Class VI myosins were first identified in *Drosophila melanogaster*⁵ and have since been found to be expressed in such diverse species as *Caenorhabditis elegans* and humans^{6–9}. As myosin VI is concentrated at actin-rich stereocilia in sensory hair cells in the inner ear it has been suggested that myosin VI is involved in the mechanical function of stereocilia^{7,10}. Consistent with this hypothesis, mutation of the myosin VI gene results in auditory malfunction¹⁰. A number of studies support the hypothesis that myosin VI is involved also in organelle transport⁹, and it may also have a role in organizing both actin filaments and microtubules¹¹. Unlike many other unconventional myosins, class VI myosin has only a single light-chain-binding site at its neck region, followed by a coiled-coil domain that would be long enough to form a two-headed structure. The presence of a large, unique insertion between the motor domain and the neck domain is particularly noteworthy. Although the mechanism by which myosin VI moves in the opposite direction on actin filaments is unknown, it was proposed that this unique insertion between the motor domain and the light-chain-binding domain of myosin VI is responsible for the reversal of the direction of movement⁴.

To elucidate the mechanism that determines the directionality of myosin movement we constructed chimaeric myosins, which were composed of the motor domain of myosin V coupled to the lever-arm/converter domain from myosin VI, which has movement in the opposite direction (Fig. 1). We also constructed a chimaeric myosin composed of the motor domain of myosin VI and the lever-arm domain of myosin V. The chimaera M5M6HMM is composed of the motor core and converter domains of myosin V (a plus-end-directed myosin) and the lever-arm domain of myosin VI (a minus-end-directed myosin). M5M6CVHMM contains the entire converter domain of myosin VI in addition to the lever-arm domain, and M5M6SHCVHMM further contains the SH1/SH2 helix of myosin VI. On the other hand, M6M5HMM is composed of the motor core and the converter domains of myosin VI (but not its large, unique insert) and the lever-arm domain of myosin V.

These chimaeric myosin heavy chains were co-expressed with calmodulin in Sf9 insect cells. The purified chimaeric myosins are composed of the heavy chain and light chains with a relative molecular mass of 17,000 (M_r 17K) (Fig. 1). The light chain changed its apparent mobility in the presence of 1 mM Ca^{2+} indicating that it is calmodulin. It was also recognized by anti-calmodulin antibodies in western blot analysis. Densitometric analysis of SDS-polyacrylamide gel electrophoresis (PAGE) revealed that the ratio of calmodulin to the heavy chain was 1.2, 0.9, 0.9 and 1.0 for M6HMM, M5M6HMM, M5M6CVHMM and M5M6SHCVHMM, respectively, as expected from one IQ motif present in the lever-arm domain of myosin VI (refs 6, 9). Furthermore, 5.7 and 5.9 mol calmodulin were bound to M5HMM and M6M5HMM, respectively, as expected from the six calmodulin-binding IQ motifs of the myosin V lever-arm domain (Fig. 1).

Both the maximal actin-activated ATPase activity and the K_{actin} (the actin concentration giving half V_{max}) of M5M6HMM were similar to M5HMM rather than M6HMM (not shown). Recent studies have revealed that the motor activity of myosin V has a

Table 1 Directionality and velocity of the movement of myosin constructs

Construct	Directionality	Velocity ($\mu\text{m s}^{-1}$)
M5HMM	Plus	0.32 ± 0.08 ($n = 112$)
M6HMM	Minus	0.29 ± 0.11 ($n = 80$)
M5M6HMM	Plus	0.13 ± 0.06 ($n = 125$)
M6M5HMM	Minus	0.05 ± 0.02 ($n = 101$)
M5M6CVHMM	Plus	0.06 ± 0.03 ($n = 98$)
M5M6SHCVHMM	Plus	0.05 ± 0.02 ($n = 91$)

Actin filament motility was observed in 25 mM KCl, 5 mM MgCl_2 , 25 mM Imidazole-HCl, pH 7.5, 1 mM EGTA, 0.5% methylcellulose; 4.5 mg ml^{-1} glucose, 216 $\mu\text{g ml}^{-1}$ glucose oxidase, 36 $\mu\text{g ml}^{-1}$ catalase and 5 mM ATP at 25 °C. All values of velocity are mean velocity $\pm$ s.d.

processive nature^{12–14}. This is reflected by its high affinity for actin in the presence of ATP¹⁴. The maximum actin-activated ATPase activity of M5M6HMM was achieved at a low actin concentration and the K_{actin} value (0.69 μM) was nearly identical to that determined for M5HMM (0.79 μM), and different from the K_{actin} determined for M6HMM (10 μM). These results indicate that the motor characteristics of M5M6HMM are inherited from myosin V but not myosin VI. These results also suggest that the motor characteristics of myosin are determined by the nature of the motor domain of the myosin molecule. A remaining question is whether or not the motor domain determines the direction of myosin movement. There is a large, unique insert present in the class VI myosin between the converter and IQ domains, and it has been proposed that this may be responsible for the change in direction of myosin movement⁴. Among the diverse members of the myosin superfamily, an insert at this position is only present in class VI myosins; therefore, if this insert is responsible for the minus-end-directed movement of myosin, it is predicted that only class VI myosin would be able to move towards the minus end of actin filaments.

To determine the direction of movement of the myosin constructs, we used F-actin filaments in the *in vitro* motility assay that

were labelled throughout with fluorescein and also with a rhodamine cap at the filament's pointed end. We developed a new method (see Methods) to produce the dual-labelled actin filaments, because the previously reported method involved the crosslinking of actin, which produced a large cap at the end of the F-actin filaments and prevented proper movement of F-actin. Our new method overcomes this problem, and both the probability of actin filament movement and the velocity of the dual-fluorescence-labelled F-actin were nearly identical to conventional, singly labelled F-actin.

The movement of these dual-fluorescence-labelled filaments of F-actin on coverslips coated with myosins can be seen under the fluorescence microscope. M5HMM moved the dual-fluorescence-labelled F-actin with the pointed end of the filaments at the front of the movement (Fig. 2). This means that myosin V moved towards the barbed end of the filaments, as is known for conventional myosins. However, M6HMM moved the F-actin filaments with the pointed end at the rear of the movement, that is; myosin VI moved towards the pointed end of F-actin filaments (Table 1). Of note, the velocities of myosin V and VI constructs in this study were 2 and 6 times faster, respectively, than those reported previously⁴. The constructs used in this study have a long coiled-coil region that are absent in the constructs used by the study in ref. 4, and it is likely that our constructs are capable of binding to the nitrocellulose-coated coverslip in a favourable conformation for motility. M5M6HMM, which contains the myosin V motor domain plus the large insertion domain of myosin VI, moved F-actin filaments with the pointed end at the front of the movement, that is, myosin moved towards the barbed end.

These results clearly show that the direction of myosin movement is solely determined by the motor domain, and that the large insert between the converter and IQ domain does not have a role in determining the direction of movement of myosin. To further confirm this conclusion, we produced M6M5HMM and examined it for its directionality. As shown in Fig. 2a, M6M5HMM (containing the myosin VI motor domain, but not the large unique insert of myosin VI) moved F-actin filaments with the barbed end at the front of the movement. This clearly indicates that the motor domain determines the direction of myosin movement and that the unique myosin VI insert is not critical for determining the directionality of myosin.

The motor domain of myosin is subdivided into a motor core domain and a converter domain. To further identify the domain critical for determining the directionality of myosin, we produced two additional chimaeric constructs, M5M6CVHMM and M5M6SHCVHMM (Fig. 1). Both myosin constructs moved F-actin filaments with the pointed end at the front of the movement, that is, myosin moved towards the barbed end. The result shows that the motor core domain specifically determines the directionality of myosin, but that neither the converter nor the lever-arm domain is critical for the direction of the movement.

The velocities of the polarity-marked actin filaments on myosin-coated cover slips are shown in Fig. 2b. Although some variation of the sliding velocity was observed, all actin filaments moved in the same direction, as determined by the myosin constructs on the coverslips. Notably, the velocities of these chimaeras are slower than either M5HMM or M6HMM. Thus, it may suggest that the converter domain does have some function in motility, although it does not determine the direction of movement. Similar reductions in velocity have been observed in chimaeric kinesin mutants¹⁵.

What components are responsible for determining the direction of myosin movement? Recent studies have suggested that the neck structure is involved in determining the directionality of the kinesin family^{15–17}, and it is thought that the interaction between the neck and the motor domain of kinesins is involved in the positioning of the second head on microtubules, thus determining the directionality. Our results clearly show that in the case of myosin the motor core domain solely determines the direction of myosin movement.

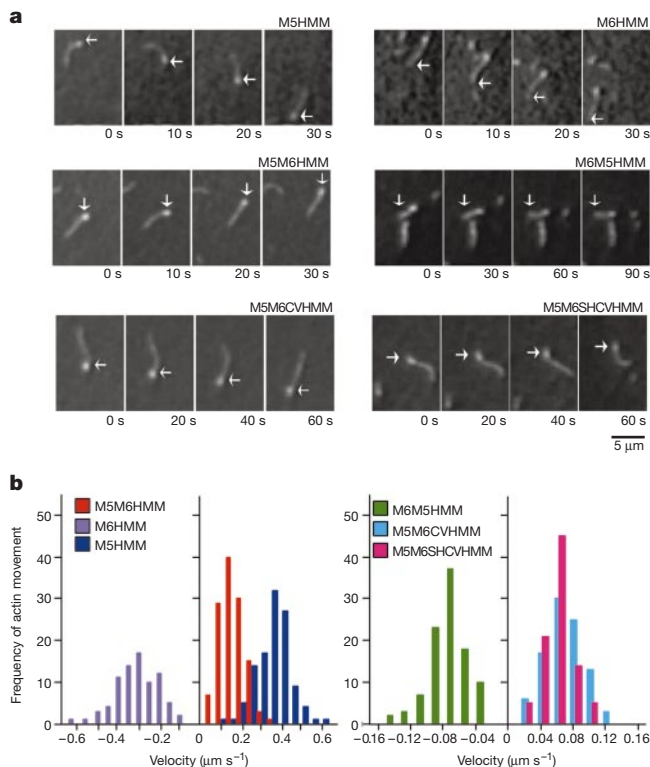


Figure 2 Direction of the movement of myosin constructs. All panels are to the same magnification. Data was obtained with a conventional *in vitro* motility assay with dual-fluorescence-labelled F-actin. **a**, Movement of dual-labelled F-actin by various myosin constructs. The bright tip on actin filaments represents the minus end of the filament. White arrows indicate the leading part of the labelled actin filaments at the front. For M5HMM the pointed end of actin filaments lead the movement, indicating that M5HMM moves towards the barbed end of F-actin. For M6HMM the barbed end of actin filaments lead the movement, indicating that M6HMM moves towards the pointed end of F-actin. Movement of F-actin by M5M6HMM, M5M6CVHMM and M5M6SHCVHMM showed that the pointed end of actin filaments lead the movement, indicating that these three constructs move towards the barbed end of F-actin. On the other hand, M6M5HMM moved F-actin filaments with the barbed end leading the movement, indicating that M6M5HMM moves towards the pointed end of F-actin. **b**, Histogram of the velocities of actin filaments having polarity markers. Movement towards the barbed end of F-actin is defined as a positive value.

During the conformational transition of the myosin head, it is suggested that the subdomains of the myosin motor domain move, and that short elements that link all subdomains (that is, the switch II and 'relay' elements) have a critical role in transmitting structural changes within the head^{18–20}. On the basis of three-dimensional structural analysis of various conformational states of the myosin head it has been proposed that the conformational change in these elements determines the positioning of the converter domain¹⁸. Therefore it is plausible that the elements connecting the subdomains of the motor domain have a role in directing the movement of myosin during the motor cycle.

The direction of myosin movement is determined by the motor core domain, not the converter domain nor the large class VI myosin-specific insert. The present finding implies that although only class VI myosin has been identified so far as a minus-end-directed myosin, it is probable that other minus-end-directed myosins are present among the other uncharacterized myosins in the myosin superfamily. Identification of the specific elements in the motor core domain that determine the directionality of myosin movement requires further biochemical, biophysical and structural studies. □

Methods

Proteins

Actin was obtained from rabbit skeletal muscle and purified²¹. Recombinant calmodulin from *Xenopus* oocytes²² was cloned and expressed in Sf9 cells as described²³.

Generation of expression vectors for myosin constructs

We produced a baculovirus transfer vector for mouse myosin VI variants in pFastBac (GibcoBRL). We obtained mouse myosin VI complementary DNA clones containing fragments 150–2565 (B10) and 1460–3708 (B2) in pBluescript. A 1679–3309 cDNA fragment of myosin VI was obtained from the B2 clone by *Kpn*–*pfl*M1 digestion and was inserted into the B10 clone after its digestion by *Kpn*1–*pfl*M1. A unique *Nhe*1 site was created at the 5' side of the initiation codon, and then the myosin VI cDNA was cut by *Nhe*1–*Kpn*1 digestion and inserted into the pFastBac baculovirus transfer vector at the polylinker region. A hexa-histidine tag sequence with a stop codon was introduced at the 3' side of the *Kpn*1 site. This construct (M6HMM) contains the entire coiled-coil domain. The mouse myosin V HMM construct (M5HMM) was produced as described²⁴.

To obtain the chimaeric construct (M5M6HMM) a unique *Spe*1 site was created at nucleotide 2266 of the myosin V cDNA and nucleotide 2299 of the myosin VI cDNA, respectively. The vectors were digested with *Spe*1, and then the myosin VI cDNA containing the sequence encoding the myosin VI unique insert, the IQ domain, the coiled-coil domain and a hexa-histidine tag was ligated into the 3' side of myosin V cDNA encoding the motor domain. The unique *Spe*1 site at the myosin V–myosin VI junction was mutated to restore the authentic amino-acid residues. For M6M5HMM unique *Nhe*1 sites were created at nucleotide 2266 of myosin V cDNA and nucleotide 2299 of myosin VI cDNA, respectively. The vectors were digested with *Nhe*1 and then the cDNA of myosin V containing the sequence encoding the IQ domain, coiled-coil domain, and hexa-histidine tag was ligated into the 3' side of myosin VI cDNA encoding the motor domain. The unique *Nhe*1 site at the myosin VI myosin–V junction was mutated to restore the authentic amino-acid sequence.

For M5M6CVHMM *Xba*1 sites were created at nucleotide 2095 and 2128 of the M5HMM and M6HMM constructs, respectively. M6HMM was digested with *Xba*1 and the fragment encoding the entire converter, the lever-arm domain and the coiled-coil domain was ligated into M5HMM digested with *Xba*1, thus eliminating the sequence encoding the myosin V converter domain, neck domain and coiled-coil domain. For M5M6SHCVHMM *Xba*1 sites were created at nucleotide 2050 and 2083 of M5HMM and M6HMM constructs, respectively, and the chimaeric construct was produced as described above. The created *Xba*1 site was mutated to restore the original amino-acid residue. The amino-acid sequence of the chimaeric constructs are as follows: M5M6HMM (Met 1–Gly 753 of myosin V plus Lys 765–Thr 1052 of myosin VI); M6M5HMM (Met 1–Gly 764 of myosin V plus Gln 754–Arg 1193 of myosin V); M5M6CVHMM (Met 1–Gly 696 of myosin V plus Phe 708–Thr 1052 of myosin VI); M5M6SHCVHMM (Met 1–Gly 685 of myosin V plus Met 697–Thr 1052 of myosin VI). Expression and purification of recombinant myosin V (M5HMM), myosin VI (M6HMM) and the chimaeric myosins (M5M6HMM, M5M6CVHMM, M5M6SHCVHMM and M6M5HMM) were carried out as described²⁴.

Preparation of the dual-labelled F-actin

F-actin (1.3 mg ml^{−1}) was first labelled with a fivefold molar excess of tetramethylrhodamine-5-(or -6)-maleimide (Molecular Probes) in the presence of 8 μM phalloidin, 25 mM KCl, 5 mM MgCl₂, 1 mM EGTA and 25 mM Imidazole-HCl, pH 7.5 (buffer A) in the dark for 40 min at 4 °C. After the reaction was stopped by adding 10 mM dithiothreitol (DTT), the fluorescently labelled F-actin was obtained by centrifugation (250,000g for 10 min). The pellet was homogenized with buffer A with 5 mM DTT and then precipitated again

(250,000g for 10 min) to wash away the residual dye. The pellet was then homogenized with buffer A with 5 mM DTT again and was diluted with the same buffer (F-actin concentration was approximately 0.03 mg ml^{−1}) and subjected to sonication to make a minus-end cap. Fragmentation of the fluorescently labelled F-actin was confirmed under the fluorescence microscope. G-actin (0.1–1 μM) was added to the fragmented, fluorescently labelled F-actin (2–15 μg ml^{−1}) in 25 mM KCl, 5 mM MgCl₂, 1 mM EGTA, 25 mM Imidazole-HCl, pH 7.5 and 10 U ml^{−1} fluorescein phalloidin (Molecular Probes). The elongation of actin filaments was then carried out overnight in the dark at 4 °C. Observation under the fluorescence microscope (Nikon Diaphoto II) and SIT camera (VE 1000 SIT, DAGE MTI) with fluorescein isothiocyanate–rhodamine dual filters (Nikon) revealed F-actin filaments with one bright cap at either end associated with a long, dim tail. Filaments that were evenly labelled without bright caps were also present. Higher concentrations of tetramethylrhodamine-labelled actin increased the probability of the capped filaments, but it also increased the number of filaments that contained more than two bright spots on each filament.

ATPase assay and *in vitro* motility assay

The steady-state ATPase activity was determined at 25 °C as described²⁵. The *in vitro* motility assay was performed as described²⁶. Actin filament velocity was calculated from the movement distance and the elapsed time in successive snapshots. We used a Student's *t*-test for statistical comparison of mean values. A value of *P* < 0.01 was considered to be significant.

Received 2 April; accepted 21 June 2001.

1. Mooseker, M. S. & Cheney, R. E. Unconventional myosins. *Annu. Rev. Cell Dev. Biol.* **11**, 633–675 (1995).
2. Titus, M. A. Unconventional myosins: new frontiers in actin-based motors. *Trends Cell Biol.* **7**, 119–123 (1997).
3. Hodge, T. & Cope, M. J. A myosin family tree. *J. Cell Sci.* **113**, 3353–3354 (2000).
4. Wells, A. L. *et al.* Myosin VI is an actin-based motor that moves backwards. *Nature* **401**, 505–508 (1999).
5. Kellerman, K. A. & Miller, K. G. An unconventional myosin heavy chain gene from *Drosophila melanogaster*. *J. Cell Biol.* **119**, 823–834 (1992).
6. Hasson, T. & Mooseker, M. S. Porcine myosin VI: characterization of a new mammalian unconventional myosin. *J. Cell Biol.* **127**, 425–440 (1994).
7. Hasson, T. *et al.* Unconventional myosin in inner ear sensory epithelia. *J. Cell Biol.* **137**, 1287–1307 (1997).
8. Baker, J. P. & Titus, M. A. A family of unconventional myosins from the nematode *Caenorhabditis elegans*. *J. Mol. Biol.* **272**, 523–535 (1997).
9. Mermall, V., McNally, J. G. & Miller, K. G. Transport of cytoplasmic particle catalyzed by an unconventional myosin in living *Drosophila* embryo. *Nature* **369**, 560–562 (1994).
10. Avraham, K. B. *et al.* The mouse Snell's walzer deafness gene encodes an unconventional myosin required for structural integrity of inner ear hair cells. *Nature Genet.* **11**, 369–375 (1995).
11. Kelleher, J. F. *et al.* Myosin VI is required for asymmetric segregation of cellular components during *C. elegans* spermatogenesis. *Curr. Biol.* **10**, 1489–1496 (2000).
12. Mehta, A. D. *et al.* Myosin-V is a processive actin-based motor. *Nature* **400**, 590–593 (1999).
13. Walker, M. L. *et al.* Two-headed binding of a processive myosin to F-actin. *Nature* **405**, 804–807 (2000).
14. Sakamoto, T., Amitani, I., Yokota, E. & Ando, T. Direct observation of processive movement by individual myosin V molecules. *Biochem. Biophys. Res. Commun.* **272**, 586–590 (2000).
15. Endow, S. A. & Waligora, K. W. Determination of kinesin motor polarity. *Science* **281**, 1200–1202 (1998).
16. Sablin, E. P. *et al.* Direction determination in the minus-end-directed kinesin motor ncd. *Nature* **395**, 813–816 (1998).
17. Endow, S. A. & Higuchi, H. A mutant of the motor protein kinesin that moves in both directions on microtubules. *Nature* **406**, 913–916 (2000).
18. Houdusse, A., Kalabokis, V. N., Himmel, D., Szent-Gyorgyi, A. G. & Cohen, C. Atomic structure of scallop myosin subfragment S1 complexed with MgADP: a novel conformation of the myosin head. *Cell* **97**, 459–470 (1999).
19. Dominguez, R., Freyzon, Y., Trybus, K. M. & Cohen, C. Crystal structure of a vertebrate smooth muscle myosin motor domain and its complex with the essential light chain: visualization of the pre-power stroke state. *Cell* **94**, 559–571 (1998).
20. Smith, C. A. & Raymont, I. X-ray structure of the magnesium(II).ADP.vanadate complex of the *Dictyostelium discoideum* myosin motor domain to 1.9 Å resolution. *Biochemistry* **35**, 5404–5417 (1996).
21. Spudis, J. A. & Watt, J. The regulation of rabbit skeletal muscle contraction. I. Biochemical studies of the interaction of the tropomyosin-troponin complex with actin and the proteolytic fragments of myosin. *J. Biol. Chem.* **246**, 4866–4871 (1971).
22. Chien, Y. H. & Dawid, I. B. Isolation and characterization of calmodulin genes from *Xenopus laevis*. *Mol. Cell. Biol.* **4**, 507–513 (1984).
23. Zhu, T., Sata, M. & Ikebe, M. Functional expression of mammalian myosin Iβ: analysis of its motor activity. *Biochemistry* **35**, 513–522 (1996).
24. Homma, K., Saito, J., Ikebe, R. & Ikebe, M. Ca²⁺ dependent regulation of the motor activity of myosin V. *J. Biol. Chem.* **275**, 34766–34771 (2000).
25. Ikebe, M. & Hartshorne, D. J. Effects of Ca²⁺ on the conformation and enzymatic activity of smooth muscle myosin. *J. Biol. Chem.* **260**, 13146–13153 (1985).
26. Sata, M., Matsuura, M. & Ikebe, M. Characterization of the motor and enzymatic properties of smooth muscle long S1 and short HMM: role of the two-headed structure on the activity and regulation of the myosin motor. *Biochemistry* **35**, 11113–11118 (1996).

Acknowledgements

We thank N. Jenkins and K. Avraham for sending us the cDNA fragments of mouse myosin V and mouse myosin VI. We also thank D. J. Schmidt for reading the manuscript.

Correspondence and requests for material should be addressed to M.I. (e-mail: mitsuo.ikebe@umassmed.edu).

Catalysis by hen egg-white lysozyme proceeds via a covalent intermediate

David J. Vocadlo*, Gideon J. Davies*†, Roger Laine‡ & Stephen G. Withers*

* Protein Engineering Network of Centres of Excellence and the Department of Chemistry, University of British Columbia, Vancouver, British Columbia V6T 1Z1, Canada

† Department of Chemistry, Structural Biology Laboratory, University of York, Heslington, York YO10 5DD, UK

‡ Departments of Biology and Chemistry, Louisiana State University, Baton Rouge, Louisiana 70803, USA

Hen egg-white lysozyme (HEWL) was the first enzyme to have its three-dimensional structure determined by X-ray diffraction techniques¹. A catalytic mechanism, featuring a long-lived oxocarbenium-ion intermediate, was proposed on the basis of model-building studies². The 'Phillips' mechanism is widely held as the paradigm for the catalytic mechanism of β -glycosidases that cleave glycosidic linkages with net retention of configuration of the anomeric centre. Studies with other retaining β -glycosidases, however, provide strong evidence pointing to a common mechanism for these enzymes that involves a covalent glycosyl-enzyme intermediate, as previously postulated³. Here we show, in three different cases using electrospray ionization mass spectrometry, a catalytically competent covalent glycosyl-enzyme intermediate during the catalytic cycle of HEWL. We also show the three-dimensional structure of this intermediate as determined by X-ray diffraction. We formulate a general catalytic mechanism for all retaining β -glycosidases that includes substrate distortion, formation of a covalent intermediate, and the electrophilic migration of C1 along the reaction coordinate.

The natural substrate of HEWL is the peptidoglycan cell wall of Gram-positive bacteria. It is composed of crosslinked oligosaccharides consisting of alternating 2-acetamido-2-deoxy-glucopyranoside (NAG) and 2-acetamido-2-deoxy-3-O-lactyl-glucopyranoside (NAM) residues. In the textbook mechanism proposed by Phillips (Fig. 1, path B), the enzyme cleaves the substrate by first binding the polysaccharide in a cleft six saccharide units long (A–F or –4 to +2). The substrate was proposed to bind in a distorted conformation in which a NAM unit in the –1 (D) subsite adopts a half-chair conformation, thereby forming a kink in the oligosaccharide between sites –1 and +1 (D and E). A catalytic acid/base, Glu 35, was suggested to protonate the glycosidic oxygen while Asp 52 stabilizes the developing oxocarbenium ion intermediate through electrostatic interactions. It was also proposed that Asp 52 acts to shield the α -face of the glycosyl oxocarbenium ion, such that it can only be intercepted by water from the β -face to yield the hydrolysed product with retained stereochemistry at C1. It was later demonstrated that the intermediate could also be intercepted by another saccharide moiety that had diffused into the active site, producing a transglycosylation product⁴.

The nature and purported stability of the ion-pair intermediate have always been points of contention. The lifetime of the glycosylation cation in water (10^{-12} s) is estimated to be just slightly longer than a bond vibration—much less than the time required for diffusion events⁵. Furthermore, there is no experimental evidence for a long-lived ion-pair intermediate at the active site of any glycosidase, including HEWL. Conversely, for every retaining β -glycosidase studied in detail to date⁶, the intermediate has been shown to be a reactive, covalent acylal ester or, in the case of retaining β -hexosaminidases using neighbouring-group participation, a covalent oxazoline intermediate⁷. The most rigorous demonstration of the covalent nature of the intermediate has been obtained by

measuring α -deuterium kinetic isotope effects (k_H/k_D , where k_H and k_D are rate constants for reactions of protium- and deuterium-labelled substrates) with substrates for which the breakdown of the intermediate is rate determining. In all of these cases, k_H/k_D has been found to be greater than unity^{6,8}. Effects greater than unity, measured on the second step, indicate that the anomeric centre undergoes rehybridization from sp^3 in the intermediate to sp^2 in the transition state. These results are consistent only with the collapse of a tetrahedral covalent intermediate through an oxocarbenium ion-like transition state. A long-lived ion pair may not be reconciled with these observations.

In order to observe the covalent intermediate experimentally, a situation in which the rate of formation (k_2) of the intermediate is significantly greater than its rate of breakdown (k_3) must be created (Fig. 1). Considerable efforts have previously been made, with HEWL, to find such a condition yet in no case have these efforts proven successful^{9,10}. This observation strongly suggests that, for HEWL, hydrolysis of the intermediate (k_3) occurs at a much greater rate than does its formation (k_2). We used a mutant enzyme in which the catalytic acid/base carboxylate residue had been replaced by the corresponding amide: the HEWL(E35Q) mutant enzyme. Mutation of the catalytic acid/base residue of HEWL and other β -retaining glycosidases has been shown to greatly reduce the rate of hydrolysis of unactivated glycosides by slowing both steps of the reaction^{11,12}. Substrates bearing activated leaving groups do not require significant acid catalytic assistance and so react at near wild-type rates to yield the glycosyl-enzyme intermediate¹². Hydrolysis of this intermediate, however, is slowed enormously as the attack of water no longer benefits from the base catalysis ordinarily provided by the enzyme. Indeed, on incubation of HEWL(E35Q) with

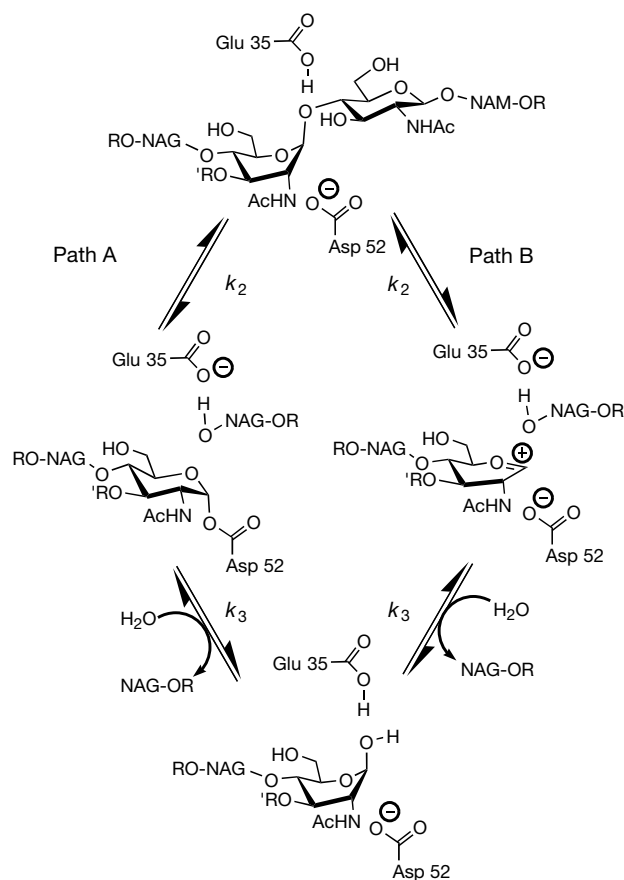


Figure 1 Two possible catalytic mechanisms for HEWL. Path A; the Koshland mechanism; path B; the Phillips mechanism. R, oligosaccharide chain; R', peptidyl side chain.

chitobiosyl fluoride (NAG₂F)¹³ we observed, using electrospray ionization mass spectrometry (ESI-MS), a mixture of the free enzyme HEWL(E35Q) and a high steady-state population (Fig. 2b) of a covalent intermediate (relative molecular mass (*M_r*) 14,719) with a mass increase (405) revealing the attachment of a chitobiosyl moiety (*M_r* 407) to HEWL. This intermediate turned over too rapidly for X-ray analysis, so a complementary strategy was sought.

To reduce the rate of the second step in a different approach, we incorporated fluorine in place of the 2-acetamido group of chitobiose, while keeping a good leaving group at the anomeric centre. Previous studies, designed to rule out a possible mechanism involving anchimeric assistance from the 2-acetamido group, demonstrated that substrates bearing hydrogen and hydroxyl substitutions at this centre are hydrolysed at similar rates to that of the parent compound bearing the acetamido group, and thus reveal that the amide functionality is not critical for catalysis^{14,15}. An electro-negative fluorine substituent at the 2 position will, however, inductively destabilize the oxocarbenium ion-like transition states leading to the formation and hydrolysis of the intermediate, slowing both steps. When a good leaving group is also incorporated to accelerate the first step, the net result is that the breakdown of the intermediate (*k₃*) is slowed to a greater extent than its formation (*k₂*), which leads to accumulation of the intermediate. Incubation of wild-type HEWL with 2-acetamido-2-deoxy-β-D-glucopyranosyl-(1→4)-2-deoxy-2-fluoro-β-D-glucopyranosyl fluoride (NAG2FGlcF; synthesis to be described elsewhere) again yielded a significant steady-state population of the covalent glycosyl-enzyme intermediate as detected by ESI-MS (Fig. 2c). The mass difference between the wild-type enzyme (*M_r* 14,316) and the intermediate (*M_r* 14,683) is 367 (368 expected), demonstrating the accumulation of the 2-fluorochitobiosyl enzyme intermediate. This intermediate was formed some 126-fold slower than with NAG₂F (*k_{cat}*/*K_m* = 9.6

± 0.1 M⁻¹ min⁻¹ and *k_{cat}*/*K_m* = 1,200 M⁻¹ min⁻¹, respectively, where *k_{cat}* is the rate constant when both substrates are at saturating concentrations and *K_m* is the Michaelis constant) but still turned over too rapidly (*k₃* > 96 min⁻¹) for X-ray crystallographic studies.

By combining these two completely independent approaches, we characterized the covalent glycosyl-enzyme intermediate formed on lysozyme by X-ray crystallography. On incubation of HEWL(E35Q) with NAG2FGlcF, the stoichiometric accumulation of a covalent intermediate could be observed by ESI-MS (Fig. 2d). The second-order rate constant governing the formation of the intermediate (*k_i*/*K_i* = 0.20 ± 0.05 M⁻¹ min⁻¹) is only 50-fold less than that observed for the wild-type enzyme interacting with the same substrate. Hydrolysis of the covalent intermediate, however, was slowed by at least 5.3 × 10⁵-fold (*k₃* < (1.8 ± 0.6) × 10⁻⁴ min⁻¹) with the result that the covalent glycosyl-enzyme intermediate was stable for several days (half-life *t*_{1/2} = 64 h) at 25 °C. The HEWL(E35Q) regenerated on such turnover could once again be labelled by incubating it with NAG2FGlcF (data not shown), indicating that the formation and breakdown of the intermediate leaves the enzyme unchanged and catalytically competent.

Diffraction data collected from crystals of the HEWL(E35Q) covalent glycosyl-enzyme formed by reaction with NAG2FGlcF reveal a complex in which the covalent linkage (~1.4 Å) between C1 and O^{δ2} of Asp 52 is unambiguous (Fig. 3a). The covalent acylal linkage is formed on the *syn* lone pair of the catalytic nucleophile as has been observed for all retaining β-glycosidases studied to date^{16–18}. One objection to the possibility of a covalent intermediate, based on the X-ray crystal structure of a NAM-NAG-NAM trisaccharide bound to the -3 to -1 (B-D) subsites of HEWL (MGM-HEWL), was that Asp52 was oriented such that only an *anti* lone pair was correctly disposed for nucleophilic attack at C1 of the -1 pyranose ring and that reorientation to align the *syn* lone pair would require major conformational change of the β-strands¹⁹. Such conformational change is not necessary and all that is observed is a rotation of 45° around χ₁ of Asp 52 such that the *syn* lone pair can engage the C1 atom to form the intermediate while leaving O^{δ1} in an unchanged position (Fig. 3b). In the covalent complex observed here, the -1 pyranose ring adopts an undistorted ⁴C₁ chair conformation, as seen in the structures of all other hexopyranose covalent intermediates so far observed (Fig. 4a)^{16–18}.

Another argument put forward against the possibility of a covalent intermediate in the HEWL reaction¹⁹ is based on the near van der Waals distance (3.2 Å) between the C1 atom of the distorted -1 (D) pyranose ring and Asp 52 observed in the MGM-HEWL product complex (Fig. 3b). These two groups cannot be expected to be significantly closer in a noncovalent complex, and in order to form the intermediate, in which the C1–O^{δ2} distance is approximately 1.45 Å, C1 and O^{δ2} have moved 1.75 Å closer. Much of this motion occurs at C1, arising from relaxation of the distorted pyranose ring to the chair conformation, leaving the sugar-enzyme contacts involving O6 and O4 almost unchanged (Fig. 3b). A conformational rearrangement of Asn 46 is necessary to prevent steric clashes between the 2-fluoro substituent and the enzyme. Although Asn 46 is disordered in the structure obtained here, it is entirely possible, with substrates in which a 2-acetamido group assumes the place of the 2-fluoro substituent, that both Asn 46 and Asp 52 may hydrogen bond to the nitrogen of the acetamido group. The ESI-MS results (see above) clearly indicate that the presence of the 2-acetamido group does not preclude formation of a covalent intermediate.

A detailed comparison of the structure determined here with that of the MGM-HEWL complex is complicated by the mobility observed in the MGM complex, manifested as high-temperature factors, that may be a consequence of it being a product complex. Fortunately, HEWL has an active-site architecture very similar to that of almost all other retaining β-glycosidases, wherein the two key carboxyl groups are positioned ~5 Å away from each other, and

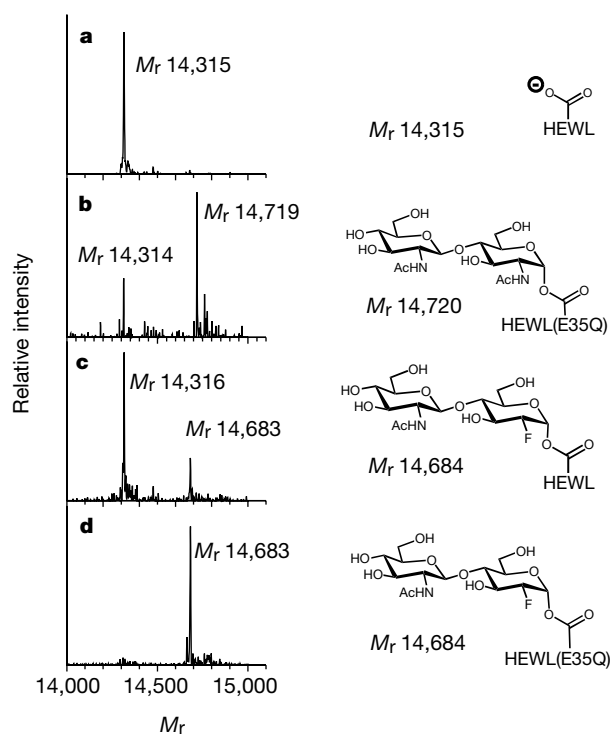


Figure 2 Transform of the ESI-MS mass spectra of HEWL and HEWL complexes. **a**, Wild-type HEWL. **b**, HEWL(E35Q) incubated with NAG2F. **c**, Wild-type HEWL incubated with NAG2FGlcF. **d**, HEWL(E35Q) incubated with NAG2FGlcF. Structures of the species corresponding to each peak observed in the mass spectra are shown to the right, with their expected relative molecular mass.

this suggests a common mechanism. Therefore, detailed insights into how HEWL accomplishes catalysis may be gained by examining the conformations of Michaelis and intermediate complexes of several β -retaining glycosidases.

Several such Michaelis complexes have now been observed using X-ray crystallography and in all of these cases the reactive pyranose adopts a skew-boat or distorted envelope conformation (Fig. 4b, c)^{18,20,21}. C1 assumes a position above the plane of the sugar ring approximately 3 Å away from the catalytic nucleophile with the

leaving group held in a pseudo-axial position by the enzyme. A comparison of representative Michaelis complexes with the covalent HEWL glycosyl enzyme structure determined in this study reveals that C1 undergoes the greatest change in position (Fig. 4b, c), whereas both the nucleophile and leaving group remain relatively fixed in space throughout the catalytic cycle. The mechanism can thus be described as an electrophilic migration of C1 from above the ring plane to below the ring plane where it easily bridges the 1.6–1.8 Å gap required to form a covalent intermediate. Additionally, the

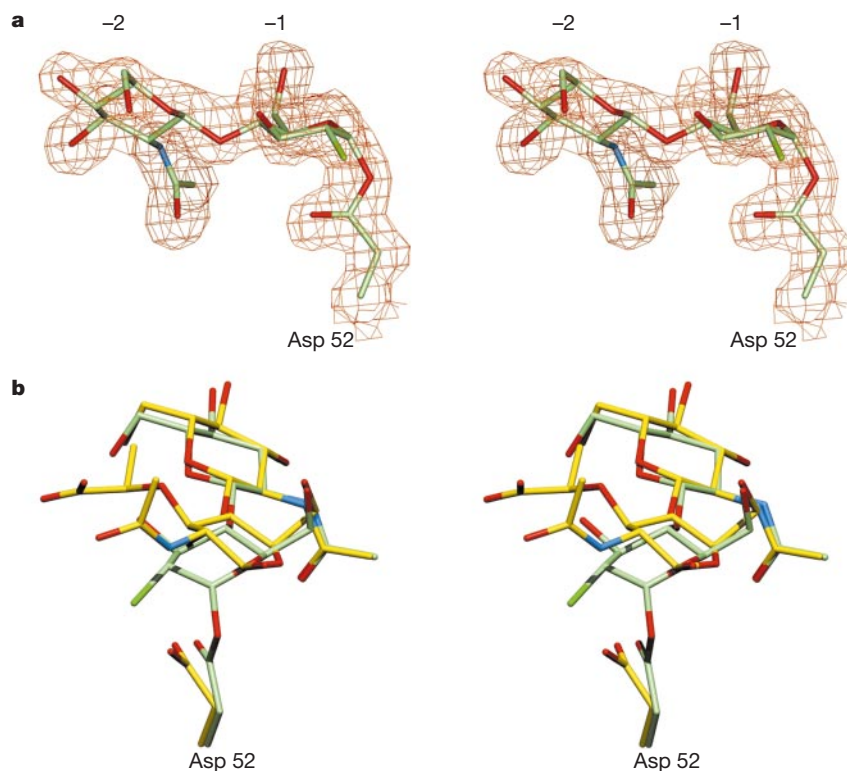


Figure 3 Structure of the covalent intermediate in the HEWL reaction. **a**, Maximum likelihood / σ_A -weighted $2F_O - F_C$ electron density for the covalent glycosyl-enzyme intermediate of HEWL, contoured at 0.4 electrons per $\text{\AA}^3$. **b**, Overlap of the HEWL covalent

intermediate and the product complex with MGM¹⁹. Formation of the intermediate is accompanied by an approximate 45° rotation around χ_1 of Asp 52.

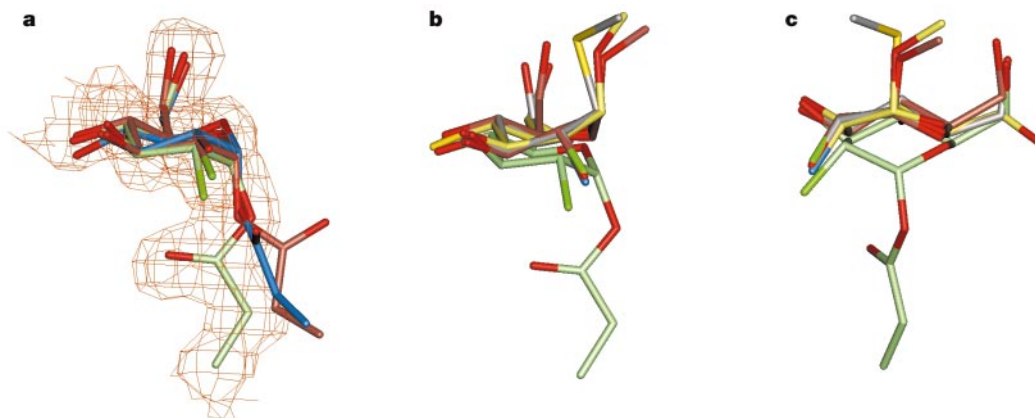


Figure 4 Comparison of hexopyranose covalent intermediates. **a**, 'Side' view of the overlap of three covalent glycosyl-enzyme intermediates: the HEWL intermediate described here (green), the cellobiosyl-enzyme intermediate from the *Cellulomonas fimi* Xyn10 (blue; ref. 17), and the 2-F cellobiosyl-enzyme intermediate from *Bacillus agaradhaerens* Cel5A (brown; ref. 18). For simplicity, only the -1 subsite sugars are shown. Electron density for the HEWL complex is shown contoured at 0.48 electrons

per $\text{\AA}^3$. **b**, **c**, 'Side' (**b**) and 'end' (**c**) view of three Michaelis complexes of retaining β -glycoside hydrolases (*Serratia marcesans* chitinase²¹ (yellow; the 2-acetamido group has been truncated at N2 for clarity), *Fusarium oxysporum* Cel7B (grey; ref. 20) and *B. agaradhaerens* Cel5A (brown; ref. 18)) together with the HEWL covalent glycosyl-enzyme intermediate (green). C1 has to migrate ~ 1.2 Å to form the covalent bond with the nucleophilic oxygen. Only the -1 subsite sugars are shown.

formation of the intermediate does not require any atoms of the α -(D) pyranose ring, other than C1, to approach Asp 52. Indeed, as C1 migrates towards the nucleophile, the 2-substituent necessarily twists away from the β -sheet of the enzyme on relaxation of the ring to the chair conformation. The transition state must be reached as C1 assumes a position along this coordinate where it approaches coplanarity with the C2, O5 and C5 atoms of the sugar ring. Kinetic isotope effects measured on the first step of catalysis by HEWL^{14,22} and other retaining glycosidases^{6,8} reveal a highly dissociative, oxocarbenium ion-like transition state that is consistent with such an electrophilic migration mechanism²³.

The hydrolysis of the covalent intermediate occurs by the near reverse of the first step, except that C1 migrates from its position below the ring plane to form a bond with a water molecule positioned in the location previously occupied by the substrate glycosidic oxygen. Water and other oxygen nucleophiles have been observed to occupy this position in several structures of covalent intermediates^{4,16,18}.

The crucially important role of Asp 52 as a nucleophile is underscored by the extremely low levels of activity that HEWL(D52N) mutants have on well defined substrates ($<0.1\%$)¹¹. The considerable residual activity ($\sim 5\%$) of the HEWL(D52A) mutant on complex cell-wall substrates has been shown to arise from chemical rescue wherein free carboxylate groups of the peptide moieties of the cell wall take the place of Asp 52 (ref. 24). Goose egg-white lysozyme (GEWL), which bears some similarity to HEWL, lacks an equivalent to Asp 52, and was long considered to provide some evidence pointing to a non-critical role for Asp 52 in HEWL. However, GEWL in fact uses a mechanism completely different from that of HEWL, which results in the net inversion of configuration at the anomeric centre²⁵.

The results of this study point to the nature of the intermediate in the HEWL-catalysed hydrolysis of glycosides. On the basis of these results, we conclude that HEWL, in common with all other retaining β -glycosidases, performs catalysis by the formation and subsequent breakdown of a covalent intermediate species and not by the formation of a long-lived ion pair. This revised mechanism unifies the substrate distortion proposed by Phillips, the covalent intermediate first postulated by Koshland, and the electrophilic migration of C1 along the reaction coordinate. It is consistent with both the anti-periplanar lone-pair hypothesis and the principle of least nuclear motion. Most importantly, this mechanism is supported by all experimental data including kinetic isotope effects, mass spectrometry, crystal structures and enzyme kinetics for all retaining β -glycosidases studied, to our knowledge. This suggests that all members of this very broad class of enzymes use the same fundamental mechanism. $\square$

Methods

ESI-MS spectra were recorded on a PE-Sciex API III mass spectrometer equipped with an ion-spray ion source. Samples were introduced by reverse-phase HPLC (C_{18} column with MeCN, H_2O and TFA elution) directly interfaced with the mass spectrometer. Crystals of the HEWL covalent intermediate were grown by the hanging-droplet vapour-diffusion method over 4 d at $16^\circ C$. The droplet, containing $2.5 \mu l$ of a solution of 7.5 mg ml^{-1} HEWL(E35Q), 20 mM NAG2FGLiF, 200 mM sodium acetate buffer at pH 5.0 and $2.5 \mu l$ of reservoir solution (200 mM sodium acetate, containing 4.0% NaCl (w/v), pH 4.5), was suspended over the reservoir solution. Crystals were transferred to cryoprotectant solution that contained 25% glycerol in addition to the reservoir buffer before freezing. Data were collected on an RAXIS IIC image-plate detector using CuK α radiation and were processed using the HKL suite of programs²⁶. All subsequent procedures used the CCP4 suite of programs unless otherwise stated²⁷. Data are 99.4% complete to $1.64 \text{ \AA}$ with an R_{merge} of 0.054 , a mean I/σ of 25 and a multiplicity of 5.2 . All structures crystallized isomorphously with the $P4_32_12$ crystal form of HEWL with approximate cell dimensions $a = b = 78.63$, $c = 36.7$, $\alpha = \beta = \gamma = 90^\circ$. The uncomplexed HEWL(E35Q) structure (not shown) was used as the starting model for refinement. Five per cent of the observations were flagged for cross-validation analysis and were 'released' through an extended simulated annealing run, to $3,000 \text{ K}$, using the program CNS²⁸. The complex structure was refined using REFMAC²⁹. The final model structure has an R_{cryst} of 0.15 , an R_{free} of 0.20 , and stereochemical deviations from target values of $0.013 \text{ \AA}$ for bonds and 1.5° for angles.

Received 8 May; accepted 28 June 2001.

- Blake, C. C. F. *et al.* Structure of hen egg-white lysozyme: A three dimensional Fourier synthesis at $2 \text{ \AA}$ resolution. *Nature* **206**, 757–761 (1965).
- Phillips, D. C. The hen egg white lysozyme molecule. *Proc. Natl Acad. Sci. USA* **57**, 484–495 (1967).
- Koshland, D. E. Stereochemistry and mechanism of enzymatic reactions. *Biol. Rev.* **28**, 416–436 (1953).
- Chipman, D. M., Pollock, J. J. & Sharon, N. Lysozyme-catalyzed hydrolysis and transglycosylation reactions of bacterial cell wall oligosaccharides. *J. Biol. Chem.* **243**, 487–496 (1968).
- Ames, T. L. & Jencks, W. P. Lifetimes of oxocarbenium ions in aqueous-solution from common ion inhibition of the solvolysis of α -azido ethers by added azide. *J. Am. Chem. Soc.* **111**, 7888–7900 (1989).
- Davies, G. J., Sinnott, M. L. & Withers, S. G. in *Comprehensive Biological Catalysis* (ed. Sinnott, M. L.) 119–208 (Academic, New York, 1998).
- Mark, B. L. *et al.* Crystallographic evidence for substrate-assisted catalysis in a bacterial β -hexosaminidase. *J. Biol. Chem.* **276**, 10330–10337 (2001).
- Sinnott, M. L. & Souchard, J. J. The mechanism of action of β -galactosidase: Effect of aglycone nature and deuterium substitution on the hydrolysis of aryl galactosides. *Biochem. J.* **133**, 89–98 (1973).
- Fink, A. L., Homer, R. & Weber, J. P. Resolution of the individual steps in the reaction of lysozyme with the trimer and hexamer of *N*-acetyl-D-glucosamine at subzero temperatures. *Biochemistry* **19**, 811–820 (1980).
- Holler, E., Rupley, J. A. & Hess, G. P. Productive and unproductive lysozyme-chitosaccharide complexes: Kinetic investigations. *Biochemistry* **14**, 2377–2385 (1975).
- Malcolm, B. A. *et al.* Site-directed mutagenesis of the catalytic residues Asp-52 and Glu-35 of chicken egg white lysozyme. *Proc. Natl Acad. Sci. USA* **86**, 133–137 (1989).
- Wang, Q., Trimbur, D., Graham, R., Warren, R. A. & Withers, S. G. Identification of the acid/base catalyst in *Agrobacterium faecalis* β -glucosidase by kinetic analysis of mutants. *Biochemistry* **34**, 14554–14562 (1995).
- Ballardie, F. W., Capon, B., Cuthbert, M. W. & Dearie, W. M. Some studies on catalysis by lysozyme. *Bioorg. Chem.* **6**, 483 (1977).
- Dahlquist, F. W., Rand-Meir, T. & Raftery, M. A. Application of secondary α -deuterium kinetic isotope effects to studies of enzyme catalysis: Glycoside hydrolysis by lysozyme and β -glucosidase. *Biochemistry* **8**, 4214–4221 (1969).
- Rand-Meir, T., Dahlquist, F. W. & Raftery, M. A. Use of synthetic substrates to study binding and catalysis by lysozyme. *Biochemistry* **8**, 4206–4214 (1969).
- Burmeister, W. P. *et al.* The crystal structures of *Sinapis alba* myrosinase and a covalent glycosyl-enzyme intermediate provide insights into the substrate recognition and active-site machinery of an S-glycosidase. *Structure* **5**, 663–675 (1997).
- Notenboom, V. *et al.* Insights into transition state stabilization of the β -1,4-glycosidase Cex by covalent intermediate accumulation in active site mutants. *Nature Struct. Biol.* **5**, 812–818 (1998).
- Davies, G. J. *et al.* Snapshots along an enzymatic reaction coordinate: Analysis of a retaining β -glycoside hydrolase. *Biochemistry* **37**, 11707–11713 (1998).
- Strynadka, N. C. J. & James, M. N. G. Lysozyme revisited: Crystallographic evidence for distortion of an *N*-acetylmuramic acid residue bound in site D. *J. Mol. Biol.* **220**, 401–424 (1991).
- Sulzenbacher, G., Driguez, H., Henrissat, B., Schulein, M. & Davies, G. J. Structure of the *Fusarium oxysporum* endoglucanase I with a nonhydrolyzable substrate analogue: Substrate distortion gives rise to the preferred axial orientation for the leaving group. *Biochemistry* **35**, 15280–15287 (1996).
- Tews, I. *et al.* Bacterial chitinase structure provides insight into catalytic mechanism and the basis of Tay-Sachs disease. *Nature Struct. Biol.* **3**, 638–648 (1996).
- Rosenberg, S. & Kirsch, J. F. Oxygen-18 leaving group kinetic isotope effects on the hydrolysis of nitrophenyl glycosides. 2. Lysozyme and β -glucosidase: acid and alkaline hydrolysis. *Biochemistry* **20**, 3196–3204 (1981).
- Federov, A. *et al.* Transition state structure of purine nucleoside phosphorylase and principles of atomic motion in enzymatic catalysis. *Biochemistry* **40**, 853–860 (2001).
- Matsumura, I., & Kirsch, J. F. Is aspartate 52 essential for catalysis by chicken egg white lysozyme? The role of natural substrate-assisted hydrolysis. *Biochemistry* **35**, 1881–1889 (1996).
- Kuroki, R., Weaver, L. H. & Matthews, B. W. Structural basis of the conversion of T4 lysozyme into a transglycosidase by reengineering the active site. *Proc. Natl Acad. Sci. USA* **96**, 8949–8954 (1999).
- Otwinski, Z. & Minor, W. in *Macromolecular Crystallography, Part A* (eds Carter, J. & Sweet, R. M.) 307–326 (Academic, London, 1997).
- Collaborative Computational Project Number 4. The CCP4 suite: Programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
- Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
- Murshodov, G. N. *et al.* Refinement of macromolecular structures by the maximum likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).

Acknowledgements

We thank N. Strynadka, Y. Luo and S. Mosimann for access to the X-ray source and computer time; J. Kirsch for supplying the DNA containing the HEWL(E35Q) clone; G. Dodson, M. Sinnott and J. Voet for critical reading of the manuscript; S. He for expert technical assistance; and H. Brummer III and S. Williams for discussions. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Protein Engineering Network of Centres of Excellence of Canada. The York Structural Biology Laboratory is supported by the Biotechnology and Biological Sciences Research Council and the Wellcome Trust. We thank the Peter Wall Institute for Advanced Studies for supporting G.J.D. under the catalytic visitor programme to the University of British Columbia. D.J.V. is a NSERC predoctoral scholar; G.J.D. is a Royal Society University Research Fellow.

Correspondence and requests for materials should be addressed to S.G.W. (e-mail: withers@chem.ubc.ca). The structure for the HEWL covalent intermediate has been deposited in the Protein DataBank under accession code 1h6m.

Try the soft option

Familiar software packages get an upgrade, and novelties get an airing.

SYSTAT 10

SPSS Science

www.spss.com/press

Graph wizard

SYSTAT 10 includes Graph Express, an application that enables users to create any SYSTAT graph from within SAS software for Windows. New statistics capabilities include mixed regression and power analysis. To create graphs, users choose from over 20 general graphics toolbar buttons that call up dialogue boxes of plot options. GraphExpress also produces immediate diagnostic plots for common analytical procedures such as factor analysis and discriminant analysis. A wizard-like Graph Gallery offers a variety of templates on which data can be plotted.

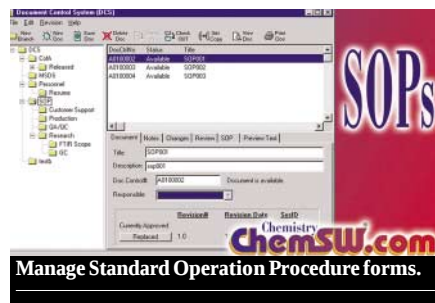
EndNote 5

ISI ResearchSoft

www.isiresearchsoft.com

Reference lists that think for themselves

Version 5 of the EndNote bibliographic software program includes a 'cite while you write' feature that can be run on Microsoft Word to locate and insert citations and produce a bibliography as the paper is being written. A travelling reference library accompanies the document, containing complete reference data for all citations, ready for colleagues to add, delete or reformat. The program also features a spell checker with multi-language and discipline-specific dictionaries, a direct link to the ISI Web of Science and a choice of over 700 output styles.



Document Control System

ChemSW

www.chemsw.com

Check it out

This Windows-based program allows laboratories to closely control documents and operating procedures while providing efficient access to the most current information. The system incorporates built-in or existing customer templates, and allows documents to be 'checked out' for revision. Secured system users can add comments to documents that are in the review and revision process. The current status of any document is available at any time, including a complete revision history.

ArrayVision

Imaging Research

www.imagingresearch.com

Latest take on analysis of gene expression arrays

This all new version of ArrayVision includes new features designed to provide true auto alignment for high-throughput fluorescent array processing and automatic export to ArrayStat, the statistical informatics program. ArrayVision 6.0 features Automation Wizards and alignment algorithms for complete walk-away batch processing. The program automatically finds the array, calculates array rotation, positions the template and aligns each spot for analysis. Additional features include simplified gene name import for clone tracking, automatic and rapid export to a variety of file formats including XML and Microsoft Excel.

Deconvolution module for TILLvisION

T.I.L.L. Photonics

www.till-photonics.de

Bring TILL images into focus

The TILL deconvolution module is an extension of the imaging software TILLvisION that can be used to remove blurring from out-of-focus and convolutional effects. A three-dimensional finite impulse response filter is used to adaptively inverse-model the point spread function that characterizes the corruption. A submicrometre-sized bead, a pseudo point light source and optical impulse function is imaged as a z stack and used in an adaptive

inverse-modelling configuration with the desired signal (an exponential that decreases rapidly) taken as an approximation to an impulse. The resulting filter is then applied to the data acquired with the microscope. The results are said to be superior to the spatial frequency techniques that introduce distortion due to frequency domain artefacts.

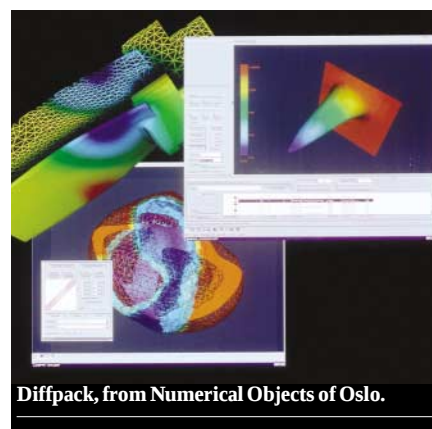
Diffpack

Numerical Objects AS

www.nobjects.com

Partial differential equations solved

Diffpack adapts object-oriented programming to numerical applications, resulting in an object-oriented numerical problem-solving program for partial differential equations (PDEs). It provides a high degree of modelling flexibility while maintaining the computational efficiency needed for real-world simulation problems in science and engineering. Diffpack allows easy modification and combination of the numerical building blocks making up an application. Suitable for traditional engineering PDE problems, it can also be used to solve non-standard problems, including pure research or problems involving a variety of mutually dependent physical effects.



Diffpack, from Numerical Objects of Oslo.

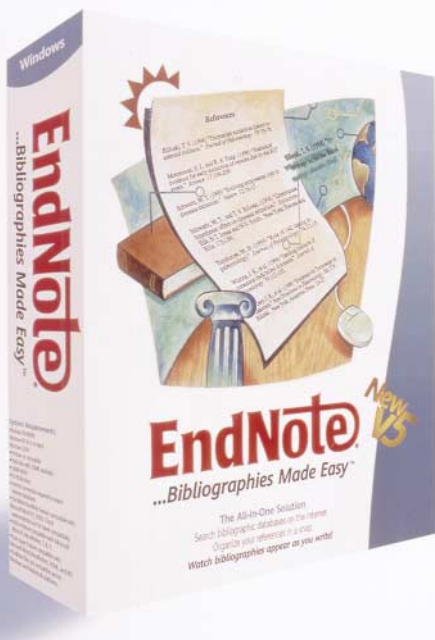
Calculus WIZ

Wolfram Research

www.wolfram.co.uk

Calculus without tears

This study aid is aimed at first-year calculus students. Based on the company's Mathematica computing system, it can be used on its own or as an add-on package to Mathematica or Mathematica for Students. Calculus WIZ can solve a broad range of calculus problems, printing the answers step by step. For example, a problem needing indefinite integration using trigonometric substitution can be entered in the program's 'solver', which will respond with the solution and paste the detailed results into a homework document, ready for printing.



Take 5: EndNote now 'cites while you write'.



AAO 1.5

Beckman Coulter www.beckmancoulter.com
AAO — for automated assay optimization

AAO 1.5, which works with the Biomek 2000 laboratory automation workstation, is designed to reduce the time required to develop fractional-factorial experiments and transform the designs into plate maps and liquid handling programs for automated pipetting. The software includes automatic data collection and analysis. Version 1.5 adds flexibility, allowing the import of a variety of more complex, multilevel designs from statistics packages such as JMP, Design Expert and Statistica, and introducing 384-well capability.

GeneSpring 4.0

a1 Biotech www.a1-biotech.co.uk
Gene expression analysis software

Genespring 4.0 enables users to locate information from the enormous banks of data generated by genomic research. An import wizard allows data to be utilized from any source and saved in HTML with full annotation. This information can then be linked to other public and proprietary data sources for fast data mining. Visualization and analysis of results can be performed and displayed in a wide range of formats. An Experimental Autoloader identifies data columns within files produced by common array technologies and allows import of multiple experimental files with a single click of the mouse.

WinLIMS 5

Quality Systems International
www.lims-software.co.uk
Mobile lab information system

This new release of WinLIMS features a Web client and a handheld client (PocketLIMS) incorporated into the standard system, along with the Windows client. Users have access to any combination of the available clients. They may remain Windows-based, advance into web operation, or take advantage of the mobility offered by PocketLIMS, which the company describes as the industry's first commercially available 'pocket client'.

Openlab

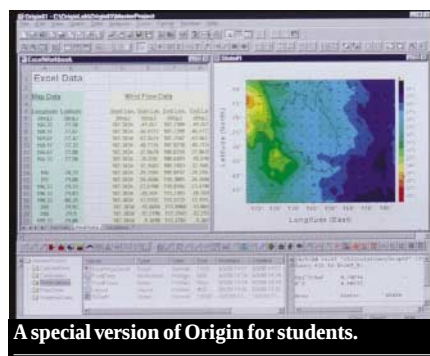
Improvizion www.improvizion.com
Cell imaging for multiple users

Openlab cell imaging software is now available in a server version for multi-user installations. Openlab Server and server modules can be installed on any number of computers with the purchase of concurrent user licences for each product. Additional licences can be purchased and authorized via e-mail.

Origin

OriginLab www.originlab.com
Student version released in US and Canada

Origin graphing and analysis software is now available in a student version to those enrolled in an accredited degree-granting institution in the USA or Canada. Origin Student Version includes all the features of standard Origin, including the ability to import many data formats, open and use Excel workbooks, customize graph attributes and create templates. Students in the USA and Canada can purchase the software through the company's website.



Maple 7

Waterloo Maple www.maplesoft.com
Mathematical computation software

The latest upgrade to the Maple suite of computational software features new mathematical algorithms in key areas such as differential equations and numerical computation. Web connectivity has been enhanced with comprehensive MathML 2.0 support, XML support and TCP/IP sockets-based access to online information. Free add-on packages and applications are also available online.

SimVector 1.00

Premier biosoft www.premierbiosoft.com
Vector drawing program

Now available for download, SimVector 1.00 supports up to 150 kb sequences and draws both circular and linear plasmids. It automates the production of graphics of vectors and provides sequence analysis to assist in the design of cloning experiments. Using GenBank web integration, a sequence can be loaded from NCBI via the Internet and displayed with annotated features automatically transformed

into graphical format. Graphics can be enhanced with patterns, styles and lines and the image can be printed, copied to the clipboard for loading into other applications, or output in web-ready graphics formats.

Microlink 751

Biodata www.microlink.co.uk
Data acquisition made easy

The Microlink 751 unit plugs into a computer's USB port. It incorporates Windmill 5 data acquisition software for Windows and can be used to measure temperature, strain, pressure, voltage and current, to monitor and control 32 digital inputs and outputs, and to count events. Eight Microlinks can be connected to one PC, and the company offers free lifetime technical support. Recalibration is automatic and can be controlled through the Windmill software. Resolution can be selected from 12 to 18 bits, allowing the user to choose high throughput or high resolution. Strain gauge bridges can be individually configured for full, half or quarter bridge wiring and normal, tensile, compressive and transverse strain can be measured.

MinWeigh

Mettler Toledo www.mt.com
Precise minimum weights

In metrology, the minimum weight is of fundamental importance. Many users carry out weighings which use only a small part of the available weighing capacity and it is in this range that the greatest inaccuracies can occur. MinWeigh is a new application, supplied with every AX, MX, and UMX analytical balance. MinWeigh allows the user to store in the balance the lowest weight which it is allowed to weigh according to his own specification and for a given measurement uncertainty. If a weight below this minimum value is weighed, MinWeigh indicates the error on the weighing terminal.

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENT

BACs • YACs • PACs cDNAs

We provide clones and colony membranes for a number of human, mouse and rat libraries as well as a custom screening service for the libraries.

ResGen™
 Invitrogen Corporation

**U.S. or Canada 800-533-4363
 FAX 256-536-9016**

Check out our web site at
www.resgen.com

Contacts

Publisher: Fabien Savenay

Editor: Paul Smaglik

Sales Director: Ben Crowe

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

Group European Manager:

Leonie Weiss (4954)

European Manager:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953), Ben Corp (4974),
Andy Douglas (4975)

Holland/ Italy: Nevin Bayoumi (4978)

Scandinavia: Silke Opstrup (4994)

Production Manager:

Billie Franklin
To send films and materials use
London address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: classified@nature.com

International

Advertising Coordinator:

Laura Pearson (4977)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Offices

France/ Belgium:

Christine Niox-Chateau
Tel + 33 (0) 1 43 20 16 51
Fax + 33 (0) 1 43 20 51 52
e-mail: c.nioxchateau@nature.com

Spain/ Portugal:

Aminda Wren
Tel + 34 91 391 3835
Fax + 34 91 319 0091
e-mail: a.wren@nature.com

Germany/ Austria/ Switzerland:

Patrick Phelan/ Kate Turner
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mails: p.phelan@nature.com
k.turner@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: classified@nature.com

US Sales Director:

Ben Crowe

US Sales Manager:

Peyton Mason

US Advertising Coordinator:

Corissa Salzman

Japan Head Office, Tokyo

Shin-Mitsuke Building (4F),
3-6 Ichigaya Tamachi Shinjuku-Ku,
Tokyo 162
Tel +81 3 3267 8751
Fax +81 3 3267 8746
e-mail: k.cowan@nature.com
Japan Manager: Kate Cowan

Inorganic identity crisis

Looking at inorganic chemistry from afar, it has the appearance of a huge monolithic entity, built primarily from large chemical companies. But in recent years, that monolith has started to shrink, as companies within the industry pursue their policies of consolidation. News of mergers and layoffs certainly isn't encouraging to the existing ranks of inorganic chemists, nor does it invite students to consider joining them.

But a closer look at basic research activity in inorganic chemistry reveals that perhaps the monolith isn't shrinking after all. Rather, its mass is being redistributed in many ways. It is being absorbed into traditionally unrelated disciplines such as proteomics; fragmented into more specialized fields such as catalysis; and rebranded into endeavours such as nanotechnology.

In the United States, there is evidence to indicate that the rising tide of life sciences may lift some fields of inorganic chemistry — particularly areas involving metallic molecules interacting with proteins (see pages 4–6). About half of the 160 recent grant proposals to the US National Science Foundation's inorganic chemistry division could be classified as bio-inorganic or organometallic research.

Elsewhere, fledgling subdisciplines could re-energize the field. For example, in Britain, basic catalysis research could spur a renaissance in some aspects of industrial chemistry (see page 8). And, in Japan and elsewhere, nanotechnology could eventually attract more investment into material sciences (see page 7).

If basic research in each of these areas is adequately funded, fuelling significant scientific progress, inorganic chemistry will once again loom large on the landscape. It will just take on the form of many smaller monuments, rather than a single large one.

Paul Smaglik

Naturejobs editor



Contents

CAREERS AND RECRUITMENT

Biology helps to redefine inorganic chemistry p4

Japan backs nanotechnology p7

Britain gets coordinated p8

WWW.NATURE.COM/ NATUREJOBS

Career centre
Information on the scientific
job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS

Inorganic chemists get cooking

Sticky problem: mussel adhesive is just one example of a useful biomaterial attracting the attentions of inorganic chemists.

The research boom in biology is helping to reshape classical chemistry disciplines, providing fresh challenges for inorganic chemists, says Steve Bunk.

If 'interdisciplinary' is the byword for today's science, then chemistry must be its exemplar. So interlaced are its various activities that the categorical relevance of chemistry's classical disciplines — organic, inorganic, physical, analytical and biochemical — is a topic of continuing debate. For inorganic chemists, this development might seem worrisome, given the huge escalation of work generated by advances in molecular biology. But even in the life sciences, collaborations now include inorganic chemists, giving them a broader employment base than ever before.

"One can say tongue-in-cheek that organic chemistry studies one element and inorganic chemistry studies the other 117," says Bruce Bursten, chemistry chair at Ohio State University. Bursten is also chair of the inorganic chemistry division in the American Chemical Society (ACS), which is further subdivided into organometallic, solid-state and bio-inorganic chemistry. At executive committee meetings, he is raising the question of whether new subcategories should be recognized by the ACS, although he declines to anticipate what they might be.

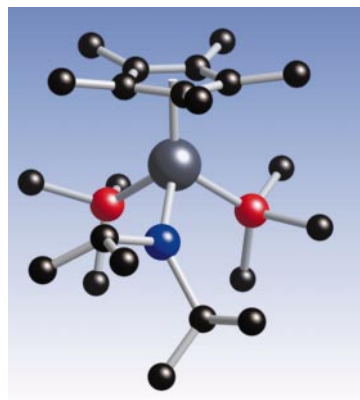
These classification issues might help to

explain otherwise confusing evidence concerning the employment outlook for inorganic chemists. On one hand, the job picture is "exuberant across all sectors of the chemical enterprise", according to *Chemical and Engineering News*, which reports "the strongest employment situation for chemists in more than a decade". Yet, when 897 job openings were posted during the annual ACS meeting in April, only 1% were for inorganic chemists. And just 4.4% of the almost 1,000 candidates who filled out forms described their work as inorganic.

Moreover, the ACS forecasts flat or only slightly increased funding for research and development in the US chemical industry over the next few years because of intensified competition from abroad and an emphasis on cost-cutting. Bulk chemicals and petroleum, bastions of inorganic chemistry, have been particularly affected by the slowdown. But leaders in inorganic chemistry suggest that the job market is buoyant in both academia and industry. One explanation is that employment is growing in areas previously unavailable to the discipline.

"DNA and RNA don't do anything without a metal ion," declares Mike Clarke, programme director for inorganic

Bruce Bursten sees inorganic chemistry as a major part of chemical studies.



chemistry at the National Science Foundation in Washington. Of about 160 funding proposals submitted to the programme last year, half involved either bio-inorganic or organometallic research. A particularly strong growth area in bio-inorganics, says Clarke, is metal-centred-protein research. In organometallics, the hot field is asymmetric catalysis, which involves changing the 'handedness' of atom arrangements in pharmaceuticals, thus altering their efficacy.

Other major research areas, says Clarke, are solid-state chemistry, new inorganic materials, polymers and magnetism. Possible applications include computer components, such as high-density storage disks. "Most of our funding is for more basic research, but more and more we tend to connect that with something that's really going to be used."

RESEARCH QUESTIONS

This gives rise to the vexed question of whether an emphasis on applied research is harming fundamental research, and what the effects might be on working chemists. At the California Institute of Technology (Caltech), chemistry and chemical engineering are in the same division, and the chemistry faculty is arranged in groups that eschew the old classifications. "The traditional disciplines still play a role in how we organize ourselves; in particular, the way we teach undergraduate students," says faculty chair David Tirrell. But research no longer conforms easily to those categories, he adds. Caltech's faculty is grouped into activities such as catalysis/synthetic methodology, materials/polymers and theoretical/computational chemistry.

Not everyone agrees with such thinking. Iowa State University's John Corbett, who specializes in inorganic solid-state chemistry, suggests that new materials design is largely a "buzz-word". For example, "Nanosphere materials are fashionable but certainly not far enough along to develop a substantial number of graduates in those fields or a substantial number of jobs," he says.

Caltech inorganic chemist John Bercaw argues that "packaging people" to take advantage of heavily funded research areas such as nanoscience does not detract from developing core strengths in the traditional disciplines. Over the years, his own interests have embraced the inorganic study of nitrogen fixation, the biochemistry of nitrogenase enzymes, polymer science and hydrocarbon oxidation. This breadth of research, which he says is typical of chemists, is a result of collaborations.

"I don't think anyone would deny that federal funding in recent years has become more targeted to initiatives," says Bursten. But to him, the concern is that other worthwhile areas could be overlooked at the expense of hot fields. As a professor, his goal in preparing students for the job market is to teach them how to solve problems, he says, no matter which employment sector they enter. Accordingly, his own research is not pointed at applications. For example, his group recently created the first compounds that directly bond a noble gas with actinide. This new

class of molecules has "no use whatsoever", he laughs.

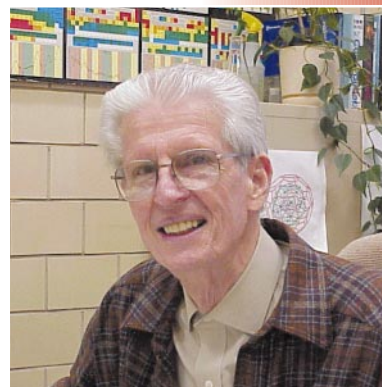
Richard Alkire, chair of chemical engineering at the University of Illinois, says that the improvement of product quality in his profession now often depends on an understanding of molecular dynamics. "Now the question is how you put together molecular systems and that's really different from digging down to find what mechanism or species is present," he says. Even so, he cautions, "People thinking about one thing all the time, even in the shower, should never become old-fashioned."

Alkire has chaired a series of chemical sciences 'round tables' in recent years, under the auspices of the National Research Council in Washington. Among his conclusions from these brainstorming sessions is that: "Universities have not worked out a mechanism for sustained interdisciplinary activities. They've worked out the birth mechanism but not the death mechanism." He means that, as yet, academia does not end its programmes well, making it difficult to transfer them to multidisciplinary centres. "The system is clicking into place, but it's probably more clunking than clicking," he says.

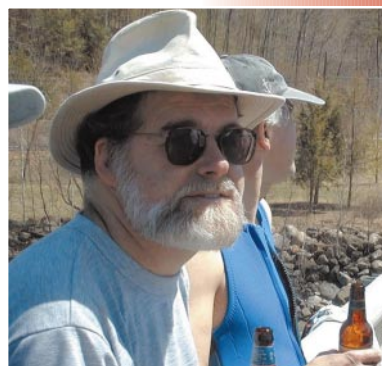
For inorganic-chemistry students, this creates the challenge of being well-prepared to enter a work culture that might emphasize organic skills. Here again, the interlaced problems of classifying disciplines and fundamental versus applied research present themselves. "It's the best of times and the worst of times in some ways, which means it's probably neither," says Richard Eisenberg, a University of Rochester professor and editor-in-chief of *Inorganic Chemistry*.

The bright side is all the new opportunities for inorganic chemists that are driven by achievements in molecular sciences, among them disease therapies based on inorganic systems or metalloenzymes, optoelectronics and new materials development. The dark side, says Eisenberg, is the effect on research. "In the way we judge proposals and particularly in the way new faculty formulate their proposals, application is not far away, which is not necessarily a good thing." The danger is that proposals can be motivated by applications that will engender no significant advance in science.

Eisenberg says the private sector once embraced the model of long-term commitment to science but quarter-to-quarter comparisons of sales now dominate R&D strategies. "I would like to see a long-term commitment to pursuing fundamental problems in industry," he says. He cites catalysts used in



John Corbett: sceptical about certain new materials.



Inorganic chemistry is a growth area, according to Mike Clarke.



Richard Eisenberg: wants long-term commitment to solving industrial problems.

industrial commodity manufacturing processes as an example, saying that they should be re-examined in the context of environmentally benign chemicals.

Jack Arrington, leader for global R&D staffing at Dow Chemical, has a different take on applied science. "There is a distinction between applied and fundamental research, to some extent, but academia does what I would call curiosity-driven research," he says. The difference is that academics often have no profit motive. He believes that in industry, applied research can lead to fundamental breakthroughs.

Alkire agrees. "I think corporations have, to a large extent, moved past this entire problem. They work on problems that are complex, they work in teams and cross these boundaries," he says. "There's no loss of touch with basic science in industry."

Arrington says: "I think in industry there has probably been a longer history of large groups working together on a common goal than in academia." This has led to a broadening of the way jobs are defined in searching for new employees, but he notes: "We probably still describe the positions in most cases along the lines of classical disciplines."

Although the company wants a scientist to carry out research in a particular area, it is expected that the employee's interests will broaden and deepen over time.

Arrington says that discussions with ACS representatives last summer indicated that the society was receiving a higher number of job-search requests from smaller companies than in the past, but he also points out that not all start-ups survive. Major companies seem to be hiring a smaller percentage of R&D people than in the past, but that reflects lower R&D budgets.

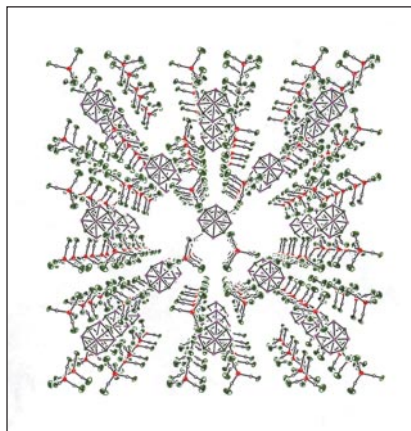
"We're hiring at about, or a little above, the rate of

attrition," he says. Until recently, organometallic chemists were in high demand at Dow for polymerization work. Those numbers are not as large now, giving way to employment of polymer chemists with some knowledge of materials engineering for electronics research.

At Ohio State, Bursten says that Dow interviewed for inorganic chemists in 2000 as usual, but he thinks the company's takeover of Union Carbide slowed down employment. The merger had been announced but was not finalized (it was completed last February), which Bursten believes "tied everybody's hands". Such consolidations of big companies do not bode well for employment, he says. "The days when most industrial chemists would get jobs at Union Carbide, Amoco, Exxon or BP are almost gone."

On the other hand, "The market for academic science is growing and will continue to grow," says Bursten. He cites solid-state chemistry as a big demand area, and Iowa State's Corbett affirms that US universities already employ up to 100 professors of solid-state chemistry. Opportunities are also growing for chemists who are willing to work in alternative fields. Jobs are proliferating in patent and

Rapid reaction: research into potential catalysts, such as this zeolite mimic, is important for industrial processes.



SHELDON G. SHORE

The lone 'alchemist' survives

Chemistry graduates who go to work for John Wasson are warned to expect the unexpected. "We tell them that with John, it's totally different from what they'll find in just about any other industrial environment," says James Hall, physics and chemistry chair at Wingate University near Charlotte, North Carolina, which supplies Wasson with his occasional helper.

Wasson says he does "bucket chemistry", Hall laughs. "And it's literal. Instead of beakers, he'll go to a supermarket and buy five-gallon buckets to use as reaction vessels." For mixing, he mounts paint stirrers on drill handles.

The company, Advanced Materials of Newhill, North Carolina, makes inorganic and organometallic chemical compounds for research and industry. Aided by his son, who is trained as a chef, and the odd graduate assistant, Wasson mostly does the work himself. In providing lab supply-houses with unusual, custom-made products, he has created hundreds of new materials every year since 1989.

"The cooking aspects of inorganic chemistry have been a mainstay of



James Hall (right) says chemists must pay their dues.

my academic and business life," he says. Trained in inorganic and physical chemistry, he did academic stints at the Universities of Kentucky and North Carolina before becoming director of chemical research for a major lithium manufacturer. In the 1980s, he co-founded two chemistry companies, one of which became a large-scale producer of lithium acetate. He says of his current business: "The biggest thrill for me is still to know I have the world's supply of a compound nobody ever made before. I can jump up and down on the desk and scream, 'You are a god!'"

Hall says he would tell any student who wanted to follow Wasson's career lead: "You've got to pay your dues, and remember how to make all the compounds you've ever made."

S.B.

environmental law, informatics and consulting, says the University of Rochester's Eisenberg.

What all this means for the education and training of inorganic-chemistry graduates is simply summarized, Corbett suggests. "The broader the better is an easy recommendation."

Steve Bunk is a freelance writer in Alexandria, Virginia.

Web links

American Chemical Society career services

♦ <http://www.acs.org/careers>

Chemical industry search engine

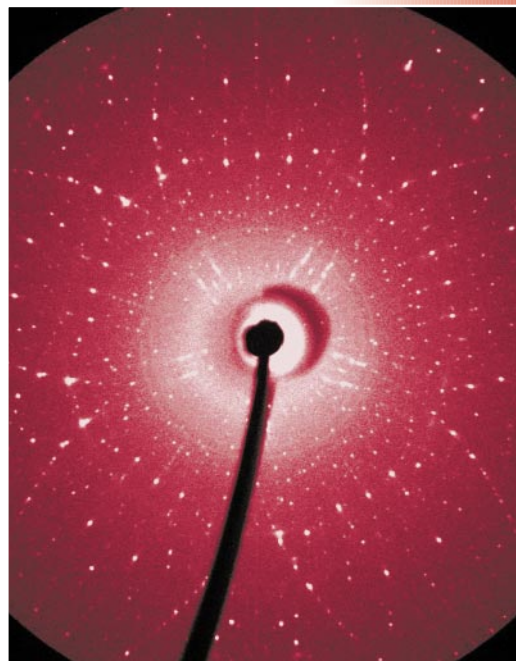
♦ <http://www.chemindustry.com>

National Academies Board on Chemical Sciences and Technology

♦ <http://www4.nationalacademies.org/cpsma/bcst.nsf>

Chemists set to gain in Japan

Renewed investment into fields such as nanotechnology should improve job prospects for inorganic chemists in Japan, says Robert Triendl.



NATIONAL INSTITUTE FOR MATERIALS SCIENCE

A decade ago, chemists in Japan were rushing into the rapidly evolving field of inorganic materials, such as ceramics. Fuelled by government-financed consortia, industrial research on inorganic materials and compounds was booming. At the same time universities and government research centres started numerous new departments and research groups.

But over the past few years, industrial research on inorganic materials has lost much of its glamour, as companies realized that associated markets were relatively small. Only a few years ago, about 20 companies were involved in research on materials such as silicon carbide, today there are only a handful, says Kazuyuki Kuroda, a chemistry professor at Waseda University in Tokyo.

Although there is general agreement that the excitement surrounding ceramics and inorganic materials has abated, research and development on inorganic materials for use in engineering and electronics continues. For Japan's troubled chemical companies — most of which have been slow to develop activities in highly value-added fields such as pharmaceuticals — such speciality chemicals often remain the only profitable part of their business.

But some observers argue that research on inorganic materials — and the whole field of inorganic chemistry — could yet see a revival. The recent discovery, by Jun Akimitsu and his colleagues at Aoyama Gakuin University, of high-temperature superconductivity in magnesium boron — a fairly standard material that can be bought off the shelf — is often cited as an example. “An unexpected discovery could soon trigger a new boom in research on inorganic materials,” says Kuroda.

In the meantime, the attention of science-funding agencies and policy-makers has shifted towards the emerging field of nanotechnology. According to one estimate, the Japanese government has spent more than US\$350 million on nanoscience and nanotechnology during 2001, a sum that compares

favourably to spending levels in the United States and is well ahead of Europe. Nanotechnology funding is set to increase further during next year's budget.

This surge in funding has already resulted in the creation of research centres and collaborations with major electronics companies that aim to develop semiconductor devices with features smaller than 100 nanometres. Also, following the decision to set up a research centre for nanomaterials at the newly created National Institute for Materials Science, many institutes and universities are planning to set up nanotechnology research centres, including the Institute for Physical and Chemical Research, Osaka University and the University of Tokyo.

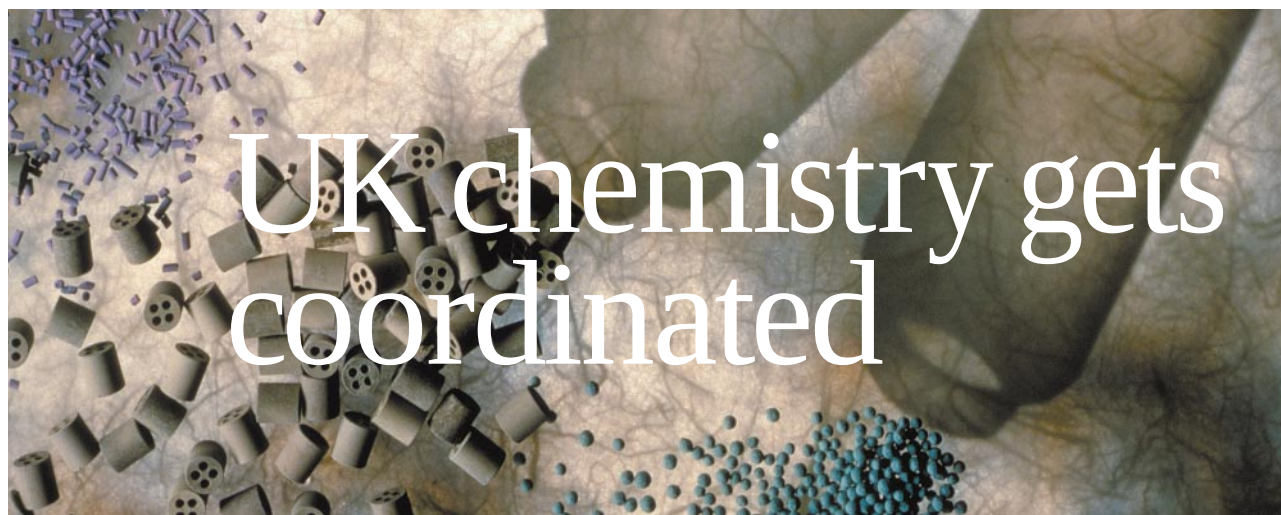
Several chemical and electronics companies — including Fujitsu, Mitsubishi Chemical and Toray Industries — have announced plans to upgrade research activities in nanotechnology. Toray revealed last month that the firm intends to spend ¥5 billion (US\$40 million) on a new nanotechnology research centre to be set up in the company's pharmaceuticals research centre in Kamakura, near Tokyo. At its completion in 2003, the centre will host between 50 and 60 full-time researchers.

The boom in nanotechnology spending is likely to fuel renewed interest in computational research and simulations, which eventually could benefit the whole field of inorganic chemistry — including basic research, which has suffered from a continuing shortage of funding.

The Japan Science and Technology Corporation, a funding body associated with the education ministry, has announced plans to launch a new initiative in computational nanoscience to start in April 2002. In addition, the Ministry of Economics and International Trade intends to support a large national computational programme called the Chemical Reaction Simulator in 2002, an activity that was started by the Japan Chemical Innovation Institute, a consortium of companies in the chemical and polymer industries.

Robert Triendl is a freelance writer based in Tokyo.

Getting a boost: inorganic chemical research could benefit from Japanese investment in nanotechnology.



Inorganic chemists face a range of possible career options when they qualify. But catalysis and biomimetics look to be reasonable bets in Britain, says Alok Jha.

The phrase 'a coordinated approach' has an extra meaning for inorganic chemists. Coordination compounds — complexes that generally feature a central metal ion to which other atoms are bound — are at the heart of many important applications. One of the most notable uses for such compounds is in catalysis.

In the United Kingdom, catalysis is an area of opportunity, according to Brian Harrison, technology director of the chemicals division at materials manufacturer Johnson Matthey. There is strong industrial interest from firms such as Matthey and Syntex, the international catalyst arm of ICI.

Harrison says that many university departments around the world are making significant advances in catalysis — research that is important for industry. For David Prest, research director at Syntex, collaborating with academia is a vital part of his company's work.

For example, Syntex has relatively extensive academic collaborations across Britain in inorganic chemistry generally and also in coordination

chemistry. These include sponsored lectureships at several universities and a professorship at the university of Glasgow. He adds that the future looks bright — the investment in universities is likely to continue as time goes on.

The next important question is who will do the work that seems to be on the horizon. "The number of people doing inorganic is small because it is daunting — you need lots of background," says Jon McLverty, professor of

chemistry at the University of Bristol.

For example, chemists who want to go into biomimetics — which involves modelling biological processes using wholly chemical systems — need to understand the biological processes that they will be emulating before they start their work. They also must be able to synthesize chemical compounds as well as learn spectroscopic techniques to analyse their work, says Sean McWhinnie, science-policy officer at the Royal Society of Chemistry.

Although individuals who have such mixed sets of skills are in short supply, employers are increasingly trying to turn that deficit into an advantage. By telling potential employees of the broad range of skills they will learn as coordination chemists, they are hoping that more of them will consider a career in the field.

The main gripe Harrison has is that UK-generated PhDs do not seem to have as well-developed problem-solving skills as their continental counterparts. McWhinnie agrees, noting that PhD students from European universities tend to have been at university for longer and are more mature.

Another problem is that the career paths for inorganic chemists can sometimes be so diverse as to be confusing. "The jobs are not in obvious places," says McLverty. "Bio-organic PhD students knew they could go into the pharmaceutical industry. People with inorganic PhDs could go into catalysis, paints, coatings, cement and analytical agencies."

For Prest, the opportunities are clear. "I expect that recruitment will never be smooth but, in general, we'll be doing steady and increasing recruitment," he says.

Malcolm Halcrow, a coordination chemist at the University of Leeds, is optimistic. As more academics come through the system and more money flows in, he expects that there will definitely be more opportunities in coordination chemistry, even if there are not many increases for chemistry as a whole.

Alok Jha is a reporter in London for *Research Fortnight*.

Optimistic: Malcolm Halcrow believes coordination chemistry could have a bright future.

